AI-NC3 INTEGRATION IN AN ADVERSARIAL CONTEXT

STRATEGIC STABILITY RISKS AND CONFIDENCE BUILDING MEASURES

ALEXA WEHSENER

ANDREW W. REDDIE

LEAH WALKER

PHILIP J. REINER

IST Institute for **SECURITY + TECHNOLOGY**

AI-NC3 Integration in an Adversarial Context:
Strategic Stability Risks and Confidence Building Measures

February 2023
Authors: Alexa Wehsener, Andrew W. Reddie, Leah Walker, and Philip J. Reiner
Design: Sophia Mauro
Footnotes: Geoffrey Ballinger

IST

About the Institute for Security and Technology

As new technologies present humanity with unprecedented capabilities, they can also pose unimagined risks to global security. The Institute for Security and Technology's (IST) mission is to bridge gaps between technology and policy leaders to help solve these emerging security problems together. Uniquely situated on the West Coast with deep ties to Washington, DC, we have the access and relationships to unite the best experts, at the right time, using the most powerful mechanisms.

Our portfolio is organized across three analytical pillars: **Innovation and Catastrophic Risk**, providing deep technical and analytical expertise on technology-derived existential threats to society; **Geopolitics of Technology**, anticipating the positive and negative security effects of emerging, disruptive technologies on the international balance of power, within states, and between governments and industries; and **Future of Digital Security**, examining the systemic security risks of societal dependence on digital technologies.

IST aims to forge crucial connections across industry, civil society, and government to solve emerging security risks before they make deleterious real-world impact. By leveraging our expertise and engaging our networks, we offer a unique problem-solving approach with a proven track record.

Acknowledgements

Table of Contents

Executive Summary

This project examines the strategic stability risks posed by integrating artificial intelligence (AI) technologies with nuclear command, control, and communications systems (NC3) across the globe. Sponsored by the U.S. Department of State's Bureau of Arms Control, Verification, and Compliance, the research aimed to clarify the often opaque vulnerabilities posed by AI technologies. Its findings underscore the value of engagement between government actors and advanced technology developers in the private sector.

Throughout the project, conversations with scientists, engineers, policymakers, and academics in both the San Francisco Bay Area and Washington, DC focused on the imperative to manage and mitigate the risks posed by AI-enabled emerging technologies. Project leaders examined the use of a suite of policy tools in the nuclear context, from unilateral AI principles and codes of conduct to multilateral consensus about the appropriate applications of AI systems. Three critical insights from this work stand out:

» **There are considerable obstacles to the creation of a fully-fledged arms control regime focused on AI technologies in general, and on the intersection of AI and nuclear capabilities specifically.** Notwithstanding existing challenges facing arms control regimes—exemplified by the collapse of the Intermediate-Range Nuclear Forces Treaty and the reticence of major players to engage in actions that might constrain technology development—uncertainties surrounding the strategic benefits and risks posed by these emerging capabilities limit the likelihood of states parties coming to the table or their ability to successfully leverage existing institutional arrangements, such as the United Nations Convention on Certain Conventional Weapons (CCW). While there are numerous conversations among academics concerning the potential regulation of compute power, data centers, data, and human capital as proxies for AI capabilities, it remains unclear whether future governance arrangements are best oriented toward the technologies themselves or the ways in which they are used (i.e., the use cases).

» **The tasks associated with the necessary risk mitigation efforts are significant.** Interviews, panel discussions, and a table-top exercise associated with this study

validated many well-established risks, from "first mover advantage" and "race to the bottom" dynamics that might proliferate the use of AI tools in a nuclear context to the "liar's dividend" enjoyed by attackers poisoning data ingested by algorithms leading to decision paralysis or inadvertent escalation. Perhaps most pronounced in our study focusing on use cases related to NC3 was the degree to which research participants discounted information presented by AI decision support tools. In the tabletop exercise, players expressed more confidence in human sources of decision making support than in machine sources. The magnitude of that gap would be a worthy area for additional study.

» **Policymakers' lack of familiarity with AI technologies prevented them from identifying ways to evaluate AI-enabled systems**. How to address the barriers to understanding AI remains a persistent question. It is unclear whether confidence in machine-based decision support is a question of familiarity. Would increasing their knowledge of machine-based decision support help them evaluate AI-enabled systems? Or, would it be more effective to supply them with access to AI expertise? That automated systems already play a significant role in intelligence, surveillance, and reconnaissance capabilities across the globe suggests that AI may eventually overcome current levels of distrust. Increased familiarity with AI may move policymakers toward informed use of these tools and systems.

Given the nascent nature of AI-NC3 integration and the uncertainty surrounding it, it is clear that an international, multi-stakeholder conversation to outline the nuclear stability risks posed by AI-based capabilities is necessary. Moreover, exercises that clarify the costs and benefits of AI-NC3 integration with engagement from both public and private sector institutions have an important role to play in these conversations, particularly given the proliferation of abstract claims in both the technical and policy fields.

Sustained strategic stability will require nuclear weapons states to share their understandings of the risks of emerging technologies across both civilian and military domains. Discussion of international guardrails could prevent accidents and inadvertent escalation. In doing so, nuclear weapons states need to think creatively about confidence building measures (CBMs) to help states mitigate risks, develop and strengthen norms, and improve decision making.

The results of this study suggest fitting CBMs into four categories:
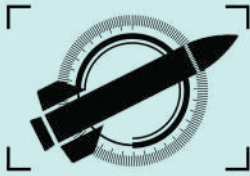
CBMs that involve agreeing to, or communicating an intent to, **renounce or limit the use of AI technologies** in certain weapon and military systems.

CBMs that encourage governments and industry players to agree on standards, guidelines, and norms related to **AI trust and safety**, as well as "responsible" use of AI technologies.
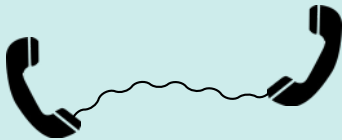
CBMs that increase **lines of communication**, such as hotlines and crisis communications links, and/or improve the quality, reliability, and security of communications in crisis.

CBMs that encourage **education and training** for policymakers, decision makers, and diplomats on sharing of AI knowledge and best practices in both the public and the private sectors.

Suggested CBMs are further elaborated in the Confidence Building Measures section of this report. The suggested CBMs are represented on a sliding scale organized by associated levels of effort and collaboration in Appendix 1: Confidence Building Measure Scales.

Introduction

During the Cold War, the U.S. government relied on nuclear physicists, engineers, and other scientific advisors to inform policy related to multilateral nonproliferation regimes and bilateral arms control agreements with the Soviet Union. In today's rapidly changing technological context, the U.S. government should likewise engage scientists and engineers with expertise in AI to understand the specific threats, vulnerabilities, and opportunities associated with integrating AI tools into NC3 systems.

But it is not sufficient to understand the technical implications of emerging and disruptive technologies. Government leaders and technologists also need to grapple with geopolitical consequences of emergent technologies and their applications—particularly when applied to high-consequence systems of systems. Their application to military capabilities and on the battlefield have the potential to alter the balance of power. And some nuclear weapons states' willingness to violate international obligations, use these technologies against civilians, and brandish the threat of nuclear conflict may increase the risks of misunderstanding and miscalculation.

Readers less familiar with nuclear systems will want to read the subsequent information on nuclear architectures and the key questions and challenges associated with the integration of AI into NC3. Readers that are more familiar with these topics can skip to the Exercise Design and Findings for novel insights from the project.

Nuclear Architectures

For decades, governments and militaries across the globe have built NC3 systems that seek to ensure positive and negative control of nuclear weapons.[1] They provide

---

1   "Positive control is the set of features and procedures that enable nuclear forces to be released when the proper authority orders it; negative control refers to the features that inhibit their use otherwise. Ideally, a nuclear state will strike a balance between positive and negative command and control, such that only an order to release nuclear weapons by the proper authority will result in a launch. But we do not live in an ideal world and new nuclear states often do not have the organizations, reliable personnel, and hardened and robust command and control architecture to implement balanced positive and negative control. Instead, they often lean one way or another. States that favor negative control are designed to more likely fail-safe — if something goes wrong, the weapons won't launch — whereas states that skew toward excessive positive control may have a tendency to fail-deadly." Vipin Narang and Ankit Panda, "Thinking Through Nuclear Command and Control in North Korea," *The Diplomat*, September 16, 2017, https://thediplomat.com/2017/09/thinking-through-nuclear-command-and-control-in-north-korea/. Narang and Panda draw on ideas put forth by Peter D. Feaver, "Command and Control in Emerging Nuclear Nations," *International Security* 17, no. 3 (Winter 1992-1993): 160-187.

the technical architecture for the collection and provision of early warning signals and intelligence, communication between leaders, and the dissemination of nuclear launch orders.[2] As computing, communication, and other technologies evolve and modernization programs advance, a rising number of scholars and policymakers have begun to consider the opportunities and risks of integrating novel AI technologies into NC3 architectures–both legacy and modernized systems.[3] However, these considerations remain mostly abstract, and while research is expanding to understand the vulnerabilities and complications these capabilities introduce into any system or system of systems, few consider the adversarial dynamics and implications created when technologies such as AI are deployed in the NC3 context. Further, while historically nuclear warheads and their means of delivery have been a substantial focus for arms control and confidence building measures in support of strategic stability and nonproliferation, comparatively little focus has been provided to the systems and networks—in digital, maritime, space, air, and terrestrial domains—upon which they rely. These challenges are exacerbated by the potential for next-generation NC3 architectures to be more closely integrated with conventional command, control, communications, computers, intelligence, surveillance and reconnaissance systems (C4ISR) contexts—further blurring the lines between nuclear and conventional weapons systems and fundamentally altering the risk landscape for inadvertent escalation.

For the past seven decades, the NC3 architectures of nuclear weapon states were developed to ensure that nuclear capabilities are always available when called upon and never used otherwise.[4] This "always/never" dilemma underpins the technical and institutional structures designed to manage the deployment and use of nuclear weapons, while also illustrating the fragility of these architectures.[5] Against this

---

2   Ashton B. Carter, "The Command and Control of Nuclear War," *Scientific American* 252, no. 1 (1985): 32-39; Feaver, "Emerging Nuclear Nations," 160-187; Daniel Shuchman, "Nuclear Strategy and the Problem of Command and Control," *Survival* 29, no. 4 (1987): 336-359; John R. Harvey, "U.S. Nuclear Command and Control for the 21st Century," *Institute for Security and Technology*, May 23, 2019: 3; Paul Bracken, *The Command and Control of Nuclear Forces*, (New Haven, CT: Yale University Press, 1983); Shaun R. Gregory, *Nuclear Command and Control in NATO: Nuclear Weapons Operations and the Strategy of Flexible Response*, (London: Palgrave Macmillan, 1996): 51-79; Robert D. Critchlow, *Nuclear Command and Control: Current Programs and Issues*, (Washington, DC: Library of Congress Congressional Research Service, 2006).

3   Thomas R. Bendel and William S. Murray, "The Bounds of the Possible: Nuclear Command and Control in the Information Age," *Comparative Strategy* 18, no. 4 (1999): 313-328; Derek Hall and Timothy Sands, "Quantum Cryptography for Nuclear Command and Control," *Computer and Information Science* 13 (January 2020): 72; Mark Fitzpatrick, "Artificial Intelligence and Nuclear Command and Control," *Survival* 61, no. 3 (2019): 81-92; Michael T. Klare, "'Skynet' Revisited: The Dangerous Allure of Nuclear Command Automation," *Arms Control Today* 50, no. 3 (2020): 10-15; James S. Johnson, "Artificial Intelligence: A Threat to Strategic Stability," *Strategic Studies Quarterly* 14, no. 1 (2020); James S. Johnson, "Artificial Intelligence & Future Warfare: Implications for International Security," *Defense & Security Analysis* 35, no. 2 (2019): 147-169.

4   "NC3 architecture" is a uniquely American concept. While the U.S. NC3 architecture is the best understood— due to the unique breadth of information publicly available—it is also likely the most complicated and most advanced. Thinking about vulnerabilities and the risk of nuclear war in less "cyber rich" contexts, therefore, is of utmost importance.

5   Feaver, "Emerging Nuclear Nations," 160-187.

backdrop, there are fears that efforts to modernize nuclear command and control systems across the globe—to include integrating AI technologies applied to signal detection and decision support—might introduce or exacerbate vulnerabilities, increasing the likelihood of accidental nuclear use, inadvertent escalation, or even intentional escalation.

For example, cyber-attacks might allow adversary operatives to penetrate nuclear or nuclear-adjacent communications systems. The most well known and widely referenced supply chain attack against nuclear related systems, Stuxnet, compromised an Iranian centrifuge in 2010. The compromise points to the soft underbelly of highly sensitive complex systems, which are still vulnerable to attack despite being air-gapped. Contractors employed by defense departments and ministries, many of which are often insecure and digitally vulnerable, are another target for adversary operatives, exemplified by the Solarwinds hack in 2020. Many legacy systems in the United States are analog, but as upgrades are made across the U.S. NC3 system of systems, new vulnerabilities will inevitably be introduced.[6] This will be the case for all nations considering modernization, especially because new systems will likely be software oriented. Even with NC3, it is likely best practice to "assume breach," as many cybersecurity experts recommend. In addition to cyberattacks that penetrate and/or interfere with NC3 networks and operations, adversaries may also use new AI techniques to spoof a nation's datasets which underpin the AI techniques or hack the data flows that feed into AI and machine learning (ML) operations.[7] These potential uses of AI may further confuse and/or degrade nuclear decision making, reducing the credibility of control and communications systems and increasing the "use-them-or-lose-them" propensity of nuclear commanders. Thus, the modernization of NC3 systems creates new risks that must be managed or overcome.

This project leveraged a table-top exercise to identify the stability risks associated with AI-NC3 integration and to provoke a discussion concerning the creation of CBMs that take into account the intersection of AI and NC3 systems across the globe. Specifically, we examined the stability risks posed by the integration of AI tools with NC3 architectures in an adversarial context. Existing analysis is overwhelmingly focused on describing the vulnerabilities associated with NC3 systems rather than considering the conditions under which these systems might be targeted. Moreover, existing analysis has yet to consider the consequences of targeting NC3 systems in an adversarial

---

6   Of note, entities such as U.S. Strategic Command would presumably create processes to mitigate these risks.

7   Ram Shankar Siva Kumar et al., "Failure Modes in Machine Learning Systems," *arXiv* preprint, (2019); Ram Shankar Siva Kumar et al., "Adversarial Machine Learning – Industry Perspectives," *arXiv* preprint, (2020).

context, beyond theorizing that this is likely to be escalatory.[8] Questions as to how escalatory and what the pathways to escalation look like remain unanswered.

To bound the study of emerging and emergent capabilities, we considered applications of machine learning and automation that are available today or likely to be deployed in the near-term. Of course, longer-term technology development may reshape the analysis—though it is worth noting that many of the applications discussed here are increasingly well-established and in some cases already integrated into a number of nuclear command and control systems across the globe.

To address this gap in the literature and to inform debates concerning appropriate measures to understand and mitigate the stability risks posed by AI-NC3 integration, this study engaged leading experts of NC3, AI, and conflict research. The study identified the escalation risks posed by the incorporation of novel technologies into NC3 architectures to drive potential confidence building and arms control measures.

Research Design

Various AI techniques have been integrated into U.S. and other NC3 systems since the Cold War. However, with new advances in both ML and other applications of AI technologies, the opportunities for application and integration have broadened and barriers have significantly decreased. This project had four aims. We first sought to understand how and why nuclear weapon states may want to integrate ML and AI into their NC3 systems. Second, we focused on understanding what repercussions and novel attack vectors would be introduced in the event of such integration efforts.

---

8   Paul Bracken, "Communication Disruption Attacks on NC3," *Institute for Security and Technology*, May 28, 2020; Shaun R. Gregory, "Command and Control of British Nuclear Weapons," *Nuclear Command and Control in NATO: Nuclear Weapons Operations and the Strategy of Flexible Response*, (London: Palgrave Macmillan, 1996): 103-129; Richard B. White, "Command and Control of India's Nuclear Forces," *The Nonproliferation Review* 21, no. 3-4 (2014): 261-274; Shaun R. Gregory, "French Nuclear Command and Control," *Nuclear Command and Control in NATO: Nuclear Weapons Operations and the Strategy of Flexible Response*, (London: Palgrave Macmillan, 1996): 130-148; Lauren J. Borja and M. V. Ramana, "Command and Control of India's Nuclear Arsenal," *Journal for Peace and Nuclear Disarmament* 3, no. 1 (May 2020): 1-20; Dmitry Adamsky, "Russian Orthodox Church and Nuclear Command and Control: A Hypothesis," *Security Studies* 28, no. 5 (2019): 1010-1039; Jeffrey Larsen, "Nuclear Command, Control, and Communications: U.S. Country Profile," *Institute for Security and Technology*, August 22, 2019; Leonid Ryabikhin, "Russia's NC3 and Early Warning Systems," *Institute for Security and Technology*, July 11, 2019; Fiona Cunningham, "Nuclear Command, Control, and Communications Systems of the People's Republic of China," *Institute for Security and Technology*, July 18, 2019; Benoît Pelopidas, "France: Nuclear Command, Control, and Communications," *Institute for Security and Technology*, June 13, 2019; Lauren J. Borja and M. V. Ramana, "Command and Control of Nuclear Weapons in India," *Institute for Security and Technology*, August 1, 2019; John Gower, "United Kingdom: Nuclear Weapons Command, Control, Communications," *Institute for Security and Technology*, August 15, 2019; Avner Cohen, "Israel's NC3 Profile: Opaque Nuclear Governance," *Institute for Security and Technology*, October 10, 2019; Feroz Hassan Khan, "Nuclear Command, Control, and Communications (NC3): The Case of Pakistan," *Institute for Security and Technology*, September 26, 2019; Myeongguk Cheon, "DPRK's NC3 System," *Institute for Security and Technology*, June 6, 2019.

Third, we aimed to understand how integrating AI and ML into NC3 may pose stability risks at different stages of nuclear escalation. Fourth, we sought to understand how CBMs could be developed and targeted to mitigate those risks, drawing from past CBM frameworks and looking to new approaches

This project brought together subject matter experts (SMEs) across industry, civil society, academia, and policy, working at the intersection of AI and NC3 for both workshop and wargame programming. Following engagement with the existing literature to evaluate NC3 and assess the current state of play vis-à-vis AI-NC3 integration, this study convened roundtable discussions, two bi-coastal workshops, and a scenario-based tabletop exercise.

Participants included senior members of the U.S. State Department, U.S. White House Office of Science and Technology Policy, and U.S. Department of Defense, as well as leading researchers from the AI and ML industries, specifically those from the AI safety and alignment community. This work was conducted under the Chatham House rule. As such, our work does not identify or attribute elements of this report to specific individuals or reveal their institutional affiliations.

AI-NC3 Integration: Use Cases and State of Play

Use Cases

There is a growing scholarly and policy-oriented literature exploring the ways in which AI technologies might be usefully employed in support of nuclear deterrence, specifically as part of NC3 modernization.[9] Some of this literature addresses problems that operators of NC3 systems already address—for example, how to analyze data

---

[9]   Peter Hayes et al., "Synthesis Report, NC3 Systems and Strategic Stability: A Global Overview," *NAPSNet Special Reports*, May 5, 2019; Peter Hayes, "Nuclear Command, Control, and Communications (NC3): Is There a Ghost in the Machine?," *Nautilus Institute*, April 9, 2018; Todd S. Sechser, Neil Narang, and Caitlin Talmadge, "Emerging Technologies and Strategic Stability in Peacetime, Crisis, and War," *Journal of Strategic Studies* 42, no. 6 (2019): 727-735; James Johnson, "The AI-Cyber Nexus: Implications for Military Escalation, Deterrence and Strategic Stability," *Journal of Cyber Policy* 4, no. 3 (2019): 442-460; Paul Scharre, "Killer Apps: The Real Dangers of an AI Arms Race," *Foreign Affairs* 98 (2019): 135; Michael C. Horowitz et al., "Strategic Competition in an Era of Artificial Intelligence," *Center for a New American Security*, July 25, 2018.

from multiple sources to make a determination regarding adversary behavior—while other challenges are perhaps more exotic. In the discussion that follows, we outline the four types of activities central to NC3 where AI technologies may be or already have been implemented to varying degrees: signal detection; decision support; target identification; and in service of implementation.

SIGNAL DETECTION, EARLY WARNING, AND SITUATIONAL AWARENESS

Signal detection is an important—if often overlooked—aspect of NC3. This is the case despite the fact that mutually assured destruction "requires nuclear-armed states to develop early-warning and agile command-and-control systems that would allow the strategic command to identify a threat and an adequate response within a limited time frame—from minutes to a few hours." Ultimately, this leads to the use of automated systems to provide early warning of adversary attack.[10] Indeed, the historical ability of automated systems to gather and analyze sensor data serves as one of the clearer examples of "AI systems" already being integrated into the NC3 mission space. Currently, these systems include radar, sonar, and infrared sensor packages deployed in fixed and mobile sensors in space, on land, and in air, as well as under and at sea. In a nuclear context, the primary role of these capabilities has traditionally been to provide data to analysts who then decide whether an adversary has launched (or is preparing to launch) a nuclear weapon—with the requirement in the United States that two systems independently make this determination.[11] While the degree to which the deployment of ML capabilities already play a significant role in early warning remains unclear across nuclear weapons states and their often technologically advanced allies, the potential applications of novel and powerful ML-driven technologies for early warning are clear.

First, the continuing advancement of ML capabilities might further improve the speed and precision of early warning. This may involve improving the types of systems doing the warning, such as the increased use of autonomous systems to surveil targets over long periods of time. Or, ML models fueled by large amounts of data could provide remote sensor systems with increased autonomy and tools to perceive, recognize, and classify adversary behavior.[12]

---

10   Vincent Boulanin et al., *Artificial Intelligence, Strategic Stability, and Nuclear Risk*, (Stockholm: SIPRI, June 2020).

11   Dual phenomenology is a U.S. policy. James M. Acton, "Escalation Through Entanglement: How the Vulnerability of Command-and-Control Systems Raises the Risks of an Inadvertent Nuclear War," *International Security* 43, no. 1 (2018): 56-99.

12   Rebecca Hersman et al., *Under the Nuclear Shadow: Situational Awareness Technology and Crisis Decisionmaking*, (Washington, DC: CSIS, March 2020).

Second, advances in machine learning analytics allow for the analysis of increasingly large, complex, disorganized, and heterogeneous data sources within and between sensor systems. Some analysts suggest that leveraging AI tools along with these diverse data sources might provide insights that a human alone would fail to recognize. At the same time, optimization methods allow models to draw inferences from smaller numbers of data points.

Third, this analysis might allow for more accurate anomaly detection when adversary behavior deviates from baseline. In real terms, this could enable predictions associated with the deployment of nuclear weapons or the preparation of military forces for an invasion of neighboring territory.[13]

Of course, the reliance upon data ingestion to feed these systems also represents a potential source of vulnerability—particularly as policymakers come to rely on intelligence products from "fused" data sources. Yet the major danger is that both the analyst and policymaker often only see the output; they are left in the dark as to what data is being used to make a determination about adversary behavior, or whether that data is manipulated to intentionally mislead the system. Emergent properties with brittleness and overfitting in advanced ML systems lend further complexity to this calculus.[14] This challenge becomes more pronounced in the context of decision support.

COMMUNICATION/DECISION SUPPORT

NC3 decision support systems exist to allow leaders to make decisions regarding the use, deployment, and movement of nuclear weapons. Decision support systems can include systems that take signal processing data and recommend posture options, targeting options, the determination of pre-selected targets and nuclear options based on simulations, and semi- to fully-automated response systems, to name a few. The Russian Perimetr system, which potentially rises to full decision automation despite being generally operated with a human in the loop, serves as an extreme example of such decision support.[15]

---

13 Alisha Anand et al., "Preemptive Discussions: The Potential Implications of Integrating Deep Learning into Early Warning Systems," *BASIC*, (2021).

14 "Brittleness occurs when any algorithm cannot generalize or adapt to conditions outside a narrow set of assumptions." M.L. Cummings, "The Surprising Brittleness of AI," *Women Corporate Directors*, January 2020; Jeff Druce et al., "Brittle AI, Causal Confusion, and Bad Mental Models: Challenges and Successes in the XAI Program," *arXiv*, June 10, 2021.

15 Boulanin et al., *Artificial Intelligence*, June 2020.

At present, AI integration already exists in nuclear systems, and, as noted above, is likely to be found within systems responsible for rapidly parsing mass amounts of information and data.[16] Currently and for the foreseeable future, AI/ML in U.S. NC3 decision support is applied in areas in which improved information processing and presentation is a value-add, rather than decision automation.[17] AI/ML decision support, in essence, is being approached as a way to get "better" information "more quickly" to the human chain of command, whether by drawing inferences that humans might not otherwise find or by replacing humans in the decision making loop. Possible integrations, specifically those of novel techniques that increase the "black box" nature of these systems, could lead to increased vulnerabilities in objective functions and throughout the learning process of the applications—not to mention the possible creation of new attack vectors for state and non-state actors to target.

TARGET IDENTIFICATION

Targeting, or the process of identifying potential targets for nuclear operations, can be viewed as a bridge between decision support and implementation. AI capabilities can already be used to discern potential targets, aid decision makers in determining the most impactful targets, and assess critical targeting decision factors such as weather, potential defensive positions, the likelihood of civilian casualties, and general target viability. Should a state have a preference for "dynamic" rather than static targeting based on well-worn doctrines, AI systems would almost certainly have a role to play in rapid retargeting.

Alongside its role in making targeting decisions, many scholars have pointed to AI systems' potential roles in the ultimate determination of nuclear use. In our estimation, the possible outsourcing of nuclear launch decisions to "dead hand" systems in the future is unclear. Although, in cases in which dead hand systems were reportedly used in the past, humans were involved in the loop.

IMPLEMENTATION

Once a decision has been made, that decision must be communicated to the delivery systems. Components of this process that are currently analog may become

---

16   Boulanin et al., *Artificial Intelligence*, June 2020.
17   Philip Reiner and Alexa Wehsener, "The Real Value of Artificial Intelligence in Nuclear Command and Control," *War on the Rocks*, November 4, 2019, https://warontherocks.com/2019/11/the-real-value-of-artificial-intelligence-in-nuclear-command-and-control/.

AI-enabled, likely those that involve transfer of data packets. This integration could reduce the amount of time it takes for a signal to pass through the system, with the inference being that automated systems will reduce communication timelines.

For each of the use cases discussed in this section, there are associated efforts to monitor, assess, and test the reliability of sensor data, communication channels, and launch controls where ML is already playing a role. It is also worth noting that there are already capabilities designed to automatically deploy absent a human in the loop. A particularly relevant example of this is missile defense interceptors.[18] Subsidiary systems within various platforms already leverage ML capabilities to perform automatic target recognition, among other tasks associated with delivering a payload to target.

Against this backdrop, NC3 and AI experts at the first San Francisco workshop highlighted several additional areas of opportunity for AI integration in NC3.

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

Experts suggested that AI-NC3 integration by multiple states threatens to introduce novel vulnerabilities, potentially shorten decision making time, and increase communication uncertainty. They also discussed the potential risks of an ever-increasing and uncritical reliance on these decision support systems. Increasing the safety and interpretability of these algorithms and discussing the criticality of safety

---

18   James Johnson, "Delegating Strategic Decision-Making to Machines: Dr. Strangelove Redux?," *Journal of Strategic Studies* 43, no. 3 (April 2022): 1-39.

and interpretability with both allies and adversaries is critical to minimizing the threat of an instant escalation incident. CBMs and other risk reduction measures regarding AI and NC3 or C3 integration will be best achieved by robust and honest conversations between states and between private and public entities concerning the risks, vulnerabilities, and potential for failure associated with these systems.

Topics discussed included the need to increase our understanding of what is actually happening within these systems, the development of tools to verify the behavior of these systems, and the need for increased deliberation time for decision makers. Participants emphasized the importance of developing AI Red Teams for any critical decision making system and urged consideration of the process required for creating a shared, international AI development and implementation framework.

The conversations also turned to the social impact of AI technologies. Today, AI can be used to attack social systems; participants suggested that, now more than ever, "compute power is translatable to political power." Political power could foreseeably degrade the command aspect of NC3, as well as increase tension and misunderstanding between nuclear weapon states, increasing nuclear risk.

Global State of Play

"We are in a strategic competition. AI will be at the center. The future of national security is at stake."

- Nuclear physicist and former National Lab participant

The United States is not the only country likely integrating AI into their NC3. This project also examined potential AI integration in NC3 in China, Russia, France, and the United Kingdom.

CHINA

As reflected in its No First Use (NFU) nuclear policy, nuclear use authority is held by top military and civilian leaders.[19] Early in its development, Chinese NC3 was reliant

---

19   Cunningham, "Nuclear Command of the People's Republic of China," July 18, 2019: 3.

upon radio-frequency communication, but since the 2000s its NC3 architecture has been upgraded to include radio communications, copper cables, fiber-optic cables, and satellites.[20] Additionally, missile brigades are equipped to use rapid communication on a closed network, known as "automated command and control," to assist in intelligence gathering for military decision makers.[21]

Alongside modernizing its existing NC3 architecture, Beijing is also expanding its strategic early warning capabilities. As articulated by Fiona Cunningham, "China might undertake distinct approaches to its NC3 relative to other nations....and as a result, the Chinese military may prove more open to leveraging certain emerging technologies, including to compensate for current shortcomings in its military capabilities."[22] Within China, technology companies have launched satellites that are capable of on-board intelligent data processing by AI-enabled chips, innovations that could transfer over into the NC3 domain to enhance early warning systems.[23] Reflecting these developments, the Chinese academic AI literature has grown rapidly. Government partnerships with top companies like Baidu and the People's Liberation Army Rocket Force (PLARF) research university are increasingly contributing to AI research conferences.[24]

There is also renewed Chinese academic interest in research on algorithms that can sense and identify data patterns. The People's Liberation Army (PLA) has made progress in remote-sensing satellite data mining and processing. It is believed to be exploring the introduction of cloud computing within its C4ISR architecture, which may lead to cloud developments within China's NC3 structure.[25] The PLA is also researching the application of quantum communication, which would enable rapid, secure communications throughout communications systems, including submarines, and could eventually be applied to NC3 architecture. Lastly, "the PLA is also actively advancing the use of AI in support of targeting and missile guidance of conventional—and potentially dual-capable or nuclear—weapons."[26]

The Chinese military will likely continue to develop and deploy NC3 systems that seek to integrate advances in AI technologies. As its capabilities progress, they will likely enhance China's launch on warning ability, including improving UAV and UUV

---

20   Cunningham, "Nuclear Command of the People's Republic of China," July 18, 2019: 5.

21   Cunningham, "Nuclear Command of the People's Republic of China," July 18, 2019: 6.

22   Cunningham, "Nuclear Command of the People's Republic of China," July 18, 2019: 2-3.

23   Saif M. Khan and Alexander Mann, "AI Chips: What They Are and Why They Matter," *Center for Security and Emerging Technology*, April 2020; Elsa Kania, "Emerging Technologies, Emerging Challenges –The Potential Employment of New Technologies in Future PLA NC3," *Institute for Security and Technology*, September 5, 2019.

24   There are increased ties between top Chinese firms, military units like the PLARF, and research universities.

25   Kania, "Emerging Technologies, Emerging Challenges," September 5, 2019, 16-17.

26   Kania, "Emerging Technologies, Emerging Challenges," September 5, 2019, 14.

capabilities and bolstering rapid response. The Chinese military, particularly since the Sino-Soviet Split, has been concerned about an incoming stealth attack going undetected, making them particularly likely to further integrate AI in early warning and defensive systems. China's NFU doctrine, as well as qualitative and quantitative characteristics of its nuclear force posture, make this integration more likely.

RUSSIA

Over the past two decades, Russia is thought to have undertaken a serious reinvestment in its NC3 systems, modernizing its nuclear capabilities and looking to restore and improve the system constructed during the Cold War.[27] Based on what can be ascertained from the outside, Russian nuclear AI integration tends to fall under three categories: early warning and ballistic missile defense, defensive posturing like Perimetr, and offensive posturing like its Poseidon and Burevestnik autonomous nuclear weapon delivery systems.[28]

The Russian military maintains a dedicated nuclear branch which controls all nuclear weapons. The system is supported by an automated system known as Combat Management Automated System (CMAS), which can be activated via two out of three nuclear briefcases held by the President, the Minister of Defense, and the Chief of General Staff.[29] The Perimetr retaliatory strike system, although secretive, allegedly remains active and a crucial part of Russian nuclear deterrence. This system is likely reliant on AI-driven software which warns of a nuclear strike based on its fusion of sensor readings.[30] Russia also claims to be developing and producing its Poseidon/ Status-6 system—an autonomous, nuclear-powered unmanned underwater vehicle capable of carrying both nuclear and conventional warheads deployed from its Oscar II-class nuclear submarines. The Russian military uses AI for robust data analysis, mainly for the purpose of early warning.[31]

Russian nuclear doctrine considers any attack on Russian NC3 a potential justification for nuclear use. The Russian military has placed an emphasis on developing redundant communication links and command posts to ensure the continuation of military operations in the aftermath of a nuclear attack. Given Russian concerns about

---

27   Ryabikhin, "Russia's NC3," July 11, 2019: 3.
28   Vincent Boulanin et al., *The Impact of Artificial Intelligence on Strategic and Nuclear Risk: Volume I, Euro-Atlantic Perspectives*, (Stockholm: SIPRI, May 2019).
29   Ryabikhin, "Russia's NC3," July 11, 2019: 3.
30   Ryabikhin, "Russia's NC3," July 11, 2019: 5.
31   Guy Faulconbridge, "Russia Produces First Set of Poseidon Super Torpedoes - Tass," *Reuters*, January 16, 2023, https://www.reuters.com/world/europe/russia-produces-first-nuclear-warheads-poseidon-super-torpedo-tass-2023-01-16/.

communications vulnerabilities (a concern, frankly, shared across the board), they may be interested in using neural networks in communication as a way to reduce susceptibility to jamming. The Russian military, like the Chinese, may also see the use of algorithms to determine least vulnerable routes and patterns for road launched missile systems as another way of using AI/ML to protect its NC3 and nuclear arsenals.

Russia, like other nuclear weapon states, has suffered false-positives in its NC3 systems in the past. The failures have also embedded a concern in the military of false-negatives, or failing to identify an incoming attack. This further justifies the need for redundant networks and command posts, and the Russian wish to ensure the survivability of mobile command posts. This fear of misperception, either a false-positive or a false-negative, drives both wariness and interest in using data to augment decision making. While the Russians value multiple streams of information, they also remain concerned about data poisoning and the corruption of intelligence via information warfare.

In the aftermath of the Russo-Ukrainian War, Russia's commercial AI sector has suffered from talent departures and lost investment and partnerships with international companies.[32] Additionally, the sector faces sanctions from the West. The Russian military and government often championed the Russian AI commercial sector and its far ranging potential, but the difficulties the sector faces, as well as the notably poor performance of Russian military systems in the Russo-Ukrainian War, casts doubt about Russia's AI capabilities, both commercial and military.

FRANCE

France's nuclear weapons are under civilian control, with the President at the top of the chain of command. Presidential decisions to use nuclear weapons are made in conjunction with the Chef d'Etat-major des Armées and the president's Chef d'Etat-major Particulier. Military officers cannot technically order nuclear launches; weapons can only be launched via civilian executive authorization. France has indicated a willingness to use nuclear retaliation against state sponsors of terrorism.[33] Communication systems, such as the RAMSES system—France's system of radio and satellite communications and radar—are well-established and protected against

---

32 "Russia's AI Disconnect: The War in Ukraine and the Looming Collapse of Russia's AI Industry," filmed April 28, 2022, *Institute for Security and Technology*, 59:49, https://www.youtube.com/watch?v=mFkN7cZnvxo.

33 Carol Ann Jones, "Counter Nuclear Command, Control, and Communications," *Institute for Security and Technology*, November 7, 2019.

attack, including reportedly being hardened and resilient to an electromagnetic pulse triggered by a nuclear detonation.[34]

Unfortunately for researchers, French law allows documents on nuclear weapons to remain classified indefinitely, meaning third parties cannot independently determine the details of French NC3 systems.[35] As a result, there is little information on the state of AI integration or intent within its NC3 systems.

UNITED KINGDOM

The United Kingdom's NC3 systems are primarily designed to ensure launch in any situation, relying on their sole use of submarine-based launch systems. This design could be due to resourcing constraints, as well as nuclear assurances from the U.S.[36] Launch authority is granted solely by the Prime Minister, although they may delegate "nuclear deputies" to authorize a launch in the event of the Prime Minister's death. In addition, Trident submarine commanders are equipped with a "dormant letter" from the Prime Minister that gives directions or delegation in case of a decapitation strike.[37] Unlike every other nuclear weapon state, the UK military has no formal role in launch authorization.[38] In the event of a crisis, since the system is submarine-based, NC3 is entirely focused on communicating orders to begin a retaliation, if necessary. Therefore, the system must be able to communicate to the submarines even in the event of a nuclear strike on the UK.[39] As the technical details of the communication systems remain highly classified, there is little information on how much AI is integrated into UK NC3. Given the UK's practice of dormant orders and pre-delegation to Trident submarine commanders, one can imagine possibilities for AI/ML integration in sensors and decision making to help facilitate the assessment of when and how to activate those orders (such as a decapitation strike on London).

AI INTEGRATION: A COMPARISON

Reflecting on the variation in technical and institutional arrangements surrounding the design of their respective NC3 architectures, it is no surprise that AI integration into

---

34   Pascale Dubois-Fernandez et al., "The ONERA RAMSES SAR System." *IEEE International Geoscience and Remote Sensing Symposium* 3 (2002): 1723-1725; Pelopidas, "France: Nuclear Command," June 13, 2019.

35   Pelopidas, "France: Nuclear Command," June 13, 2019: 3-4.

36   Gower, "United Kingdom: Nuclear Weapon Command," September 12, 2019.

37   Gower, "United Kingdom: Nuclear Weapon Command," September 12, 2019: 6.

38   Gower, "United Kingdom: Nuclear Weapon Command," September 12, 2019: 6.

39   Gower, "United Kingdom: Nuclear Weapon Command," September 12, 2019: 8.

international NC3 systems differs from the United States. To date, these differences are most salient in the Chinese and Russian cases. As a result, American decision making and strategy concerning NC3 should not be conflated with that of our allies or adversaries. Such comparisons are unfortunately common, given that the best open source information that scholars and analysts have concerning NC3 architectures relates to the U.S. case.

Ideally, Chinese, and Russian perspectives would be included in discussions of AI and NC3 failures and optimal use, but Moscow and Beijing have hitherto largely avoided coming to the table to discuss the risks to nuclear deterrence posed by emerging technologies writ large. In both states, however, AI integration has ostensibly been a part of broader efforts to modernize their respective nuclear forces. China's push to develop AI appears driven by concern over being caught off guard by its adversaries' conventional, nuclear, and AI capabilities. In addition, its focus on AI development also reflects its aim to continue to make advancements to its own NC3 and conventional C2 systems. A particularly salient element of our discussions surrounds the continuation of China's no-first use policy. Beijing's "intelligentization" also continues to reflect broader concerns regarding centralization, as does its leveraging of "civil-military fusion," though the degree to which this fusion has been successful remains open to debate. More specifically, there are indications that the PLA is looking to integrate AI into C4ISR.[40] Russia, on the other hand, continues to appear most likely to implement AI systems to guarantee its survivability in a manner reminiscent of earlier Soviet-era systems.[41] However, there does appear to be fear in Moscow in regards to a potential "first mover advantage," in which the first mover in this space might accrue overwhelming advantages. In both cases, hedging appears to play a significant role in driving AI-NC3 integration.

In Europe, the smaller arsenals of the United Kingdom and France, proximate external security threats, and domestic political contexts drive discussions concerning nuclear modernization and AI-NC3 integration. In France, the relationship between civilian authorization for nuclear use and the empowerment of military actors is uncertain, reflecting the need to command these systems both in peacetime and wartime.

In all nuclear states, it is worth noting an overriding concern regarding "jumping to AI/ deep learning solutions" where they might not otherwise be necessary or appropriate,

---

40 Elsa Kania, "Battlefield Singularity: China's Rise in Artificial Intelligence and Future Military Capabilities," *Center for a New American Security*, November 28, 2017.

41 Newer Russian weaponry like Poseidon, Burevestnik, and Zircon have the same aim as Perimetr: ensuring a second strike capability in the face of U.S. missile defense.
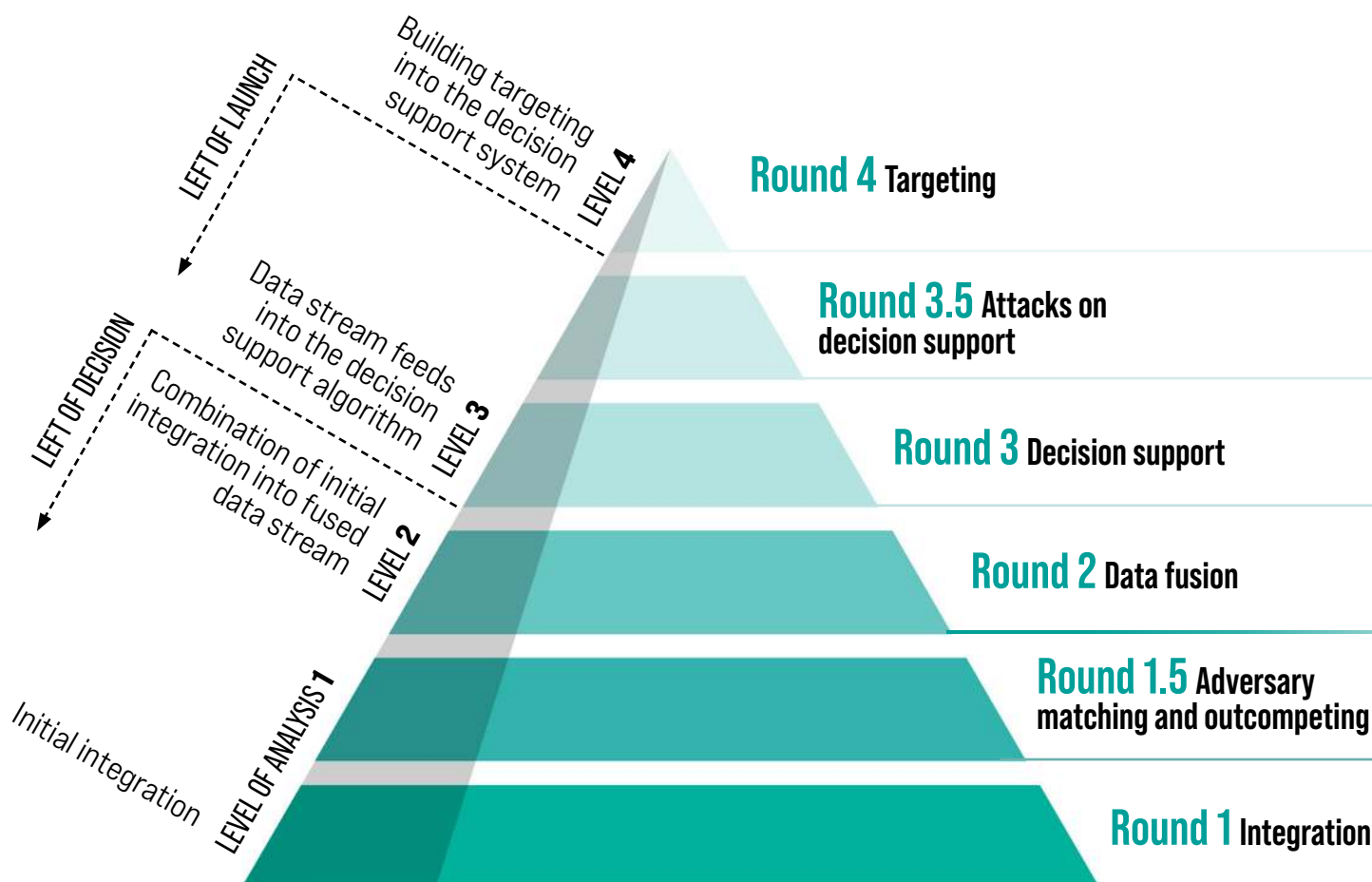
whether motivated by external security threats, bureaucratic politics, or simply profit driven defense contracting solutions. It remains to be seen whether these concerns are overcome prior to the actual deployment of advanced integrated systems or the creation of programs of record.

Exercise Design and Findings

To make salient some of the stability risks discussed in the academic literature pertaining to AI technologies, our BOUNDAR(ies) table-top exercise (TTX) asked participants to serve as national security advisors to an executive and engage with a hypothetical crisis involving four fictional countries. Each scenario engaged with different levels of conflict escalation while also presenting a series of different questions concerning AI-NC3 integration. The research questions that underpinned each scenario are included below and the full design documentation can be found in the Appendix. Rather than engaging participants in an abstract or general discussion of policy-making challenges, the TTX sought to immerse players in a decision making context in which there were stakes associated with their decisions.

As discussed, the research included technical AI researchers at every stage of the project. Their inclusion proved especially important during the exercise since they knew the right questions to ask. If the exercise had consisted of a room full of decision makers with limited technical AI backgrounds, the results likely would have been very different. The exercise successfully elicited important technical questions from the AI researchers, simultaneously underscoring the need to ensure decision makers understand these issues.

The below chart was used in the initial design of the game in order to determine the levels of integration that would be built throughout the game.

LEFT OF LAUNCH

Building targeting into the decision support system — LEVEL 4

Data stream feeds into the decision support algorithm — LEVEL 3

LEFT OF DECISION

Combination of initial integration into fused data stream — LEVEL 2

Initial integration — LEVEL OF ANALYSIS 1

Round 4 Targeting

Round 3.5 Attacks on decision support

Round 3 Decision support

Round 2 Data fusion

Round 1.5 Adversary matching and outcompeting

Round 1 Integration

Importantly, in our view, we engage both with decisions to develop and deploy AI systems in a nuclear context as well as with the players' responses to various "use" scenarios in which the presence of AI systems are theorized to drive instability.

EXERCISE FINDINGS

The scenarios provided fodder for rich discussion among participants. Below, we have identified the concerns raised that we believe are most salient. Importantly, many of the participants' thoughts reflected technical concerns as well as the impact of the strategic context in which these systems were deployed.

Measurement, Accuracy, and Transparency

TTX participants were concerned about the reliability and transparency of the AI systems described in the scenarios, particularly surrounding their use in decision support. Participants raised concerns about algorithmic accuracy, the lack of transparency on military system testing and evaluation, and the lack of comprehensive accuracy statistics.

The exercise also revealed the extent to which technical researchers, decision makers, and policymakers use the same words to mean very different things in ways that can generate confusion and ambiguity. The AI researchers questioned whether the AI-enabled systems described in the exercise were realistic. Policymakers, meanwhile, expressed discomfort with and distrust in the same systems, unrelated to how technically realistic they were. Rather, they tended to associate their distrust of systems with those that are hard to understand. We observed that while technical participants worried about the fidelity to reality, policy experts did not know the difference.

Explainable AI (XAI) is a technical tool or research area that seeks to provide "a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms."[42] XAI techniques can be integrated across an ML model to provide increased transparency into opaque systems, thus enabling AI explainability.[43] However, existing XAI methods are limited and additional research is needed to improve the integration of explainability into AI model development. Policy makers often use the term 'explainability' to refer to subject matter experts' ability to describe to them how the model arrived at its conclusions. The slight definitional nuance between building trust by implementing a technical method built into the AI model, a technical AI research area and education between SME and policymaker is large enough to create vastly different understandings that would collide in a nuclear use scenario in which AI-enabled NC3 systems are heavily relied upon by decision makers.

It is natural that decision makers will have fundamental unfamiliarity with the technical aspects of AI-enabled systems. This unfamiliarity makes it all the more crucial to identify what types of information decision makers need to be comfortable with in order to place a level of trust in AI-enabled systems.

---

42   "Explainable AI (XAI)," *IBM*, 2023, https://www.ibm.com/watson/explainable-ai.

43   "Methods exist for analyzing the data used to develop models (pre-modeling), incorporating interpretability into the architecture of a system (explainable modeling), and producing post-hoc explanations of system behavior (post-modeling)." Violet Turri, "What is Explainable AI?," C*arnegie Mellon University Software Engineering Institute*, January 17, 2022, https://insights.sei.cmu.edu/blog/what-is-explainable-ai/.

Participants also referenced common human-computer interaction concerns–not least that decision makers might overestimate the accuracy of system outputs ("automation bias"), especially relative to traditional intelligence information presented by human operators. The discussion also touched on the human bias problem in the translation of probabilistic intelligence into policy. Further research could expand on these baseline insights.

The breadth of technical and policy specialists who participated in this project highlighted not only the difference in expertise, but the difference in how the two groups think about these challenges. It also provides valuable insight about what each group needs to know in order to advance nuclear safety and decision support from the positions they occupy.

AI Safety and Assurance

Testing, evaluation, verification, and validation (TEVV) represented a top priority for the TTX participants for each of the systems that they were presented with. Participants frequently mentioned explainability; data sources/inputs and related data poisoning concerns; characterizing "expected" vs. "unexpected" operation of algorithms; and the robustness of the algorithms as being central to their willingness to trust the information provided by AI systems.

AI assurance, safety, and security loomed large. The same algorithm, depending on the use scenario, can have very different safety concerns—safety is not just an abstract concept attached to the algorithm. Indeed, AI systems should not only be tested for reliability, but also for interpretability and explainability.

Adversary Mirroring and Competing

Participants in the exercise quickly became concerned about adversary use of AI systems. These concerns included worries about the comparable or elevated levels of bias in adversary systems and the inefficiency or inability to mitigate that bias. There was also concern about the other sides' TEVV or lack thereof.

Participants worried that as AI technologies mature and states try to rapidly develop AI capabilities, adversaries might be willing to bypass testing for reliability in the rush to get systems online. There are reasons to be concerned that bureaucratic politics might make this risk more salient. Participants also raised that there are both advantages and disadvantages to encouraging adversaries to engage in a thorough testing of capabilities. Better testing and evaluation practices creates better systems–and

when it comes to NC3, the more reliable the system, the lower the risk of nuclear war. Meanwhile, providing technical testing and evaluation (T&E) tools could be used as a CBM for standardization, enabling states to 'speak in the same language' in terms of measurements. Sharing T&E procedures to an extent could also be unhelpful, if too much is shared, adversaries stand to gain intelligence to better enable their own systems. Walking the information sharing tightrope is a constant challenge.

Intelligence: Human vs. Machine

During the TTX, participants engaged in lengthy discussions about whether bias concerns were more pronounced for algorithms compared to data flows managed and gathered by human operators. Some participants raised concerns that decision makers are already trained to calculate a level of doubt and accuracy with human intelligence, but do not have a calibrated value of trust and doubt when it comes to information derived from algorithms and systems of systems driven by AI-powered decision support processes. Participants noted that while human intelligence can be explained and rationalized, machines will not be able to explain their process for arriving at a certain outcome. However, several participants cautioned that perhaps there is a tendency to overestimate the accuracy of human intelligence.

Redistributing R&D

Interestingly, some participants throughout the game showed a willingness to reallocate resources from AI research and development (R&D) towards advancing and strengthening conventional forces and nuclear weapons delivery systems. Participants also suggested that the systems in the game be treated as demos, rather than fully deployed systems, and even proposed signaling to the adversary that the systems were still in the research and development stage.

Cyber Threats and Information Flows

During the game, there were major fears associated with data poisoning. Participants often questioned how they could be certain that their systems had not fallen prey to a data poisoning attack. This fear became particularly pronounced during situations in which policymakers are relying on singular systems that leverage fused data streams. There were also concerns about broader cyber attacks and information operations. In particular, the lack of credibility in the system might lead policymakers to mistrust it when correct, and trust it when incorrect. Participants thus discussed the need to verify the credibility of information. Some participants encouraged working back to the last

trusted datasets in cases of a suspected breach and raised associated questions as to whether this was technically feasible.

Alternative Applications

There was uneven support for targeting tools posed within the TTX: "CasX" (a casualty mitigation algorithm) was viewed as fairly unproblematic while "DialY" (a risk of escalation analysis algorithm) was far more controversial. There was some discussion that DialY might be more appropriate in a non-nuclear, operational/tactical context or would be useful if tweaked to provide options for reducing risk.

While the participants were wary of many of the systems, there was still a willingness to keep the systems 'on' during a crisis, and to use the crisis as a way to test and measure the outputs of the system to obtain better metrics of accuracy and reliability. Participants called for clarity on how exactly an algorithm would be 'used' in the scenario. A technical expert astutely raised that this specificity is important. For instance, if the algorithm was used for hypothesis generation, suggestions (of attack) were treated as "ideas" that need verification. Conversely, if the algorithm was used as an hypothesis verification system, then the predictions were treated as verified. Technical participants generally felt that the former is more defensible and that it underlined why interpretability is important. In other words, they recommended using interpretability as a start of the hypothesis verification process. The same algorithm, depending on the use scenario, can have very different safety concerns. This cements the need to be mindful that safety is not merely an abstract concept attached to the algorithm.

Some participants were willing to test AI systems during a crisis, but were not comfortable with using them. It remains to be seen whether the passage of time and comfort with the use of these tools in lower stakes settings might alter this calculation.

Finally, and as we had hoped, throughout the exercise participants continually raised processes, policies, and procedures that build confidence in AI systems between and among government decision makers.

Improving AI Safety: Technical Solutions

The primary avenue for addressing the concerns associated with AI-NC3 integration is to continue bridging the gap between Silicon Valley and Washington, DC. In these interactions, participants noted the importance of engaging in a "red team mentality" as technical challenges are presented (and overcome) and as escalation risks associated with the development and deployment of these systems are managed. This focus on system safety also includes implementing security measures at every step, from the hardware supply chain to data sourcing to the operator, to create the technical foundation for AI "trustworthiness."

Participants also discussed the imperative that the United States continue working to bring China and Russia to the table in developing a framework for safe AI development and implementation. Measures for verifying compliance with developed frameworks will be needed to improve trust between states. This might include zero-knowledge reviews of AI algorithms and tracking development and purchase of compute power, using both licenses for private computing firms and checks on the destinations of physical processors.[44]

To avoid the destabilizing effect of instant response launches due to automation, participants noted that it is paramount that AI-integrated NC3 systems be able to expand the time-horizon available to decision makers, giving space to check for false alarms and purposeful deception before making an irreversible decision.

> *"Stability comes from confidence, uncertainty decreases stability"*
>
> - Former senior U.S. government participant

---

44   Nitin Singh, Pankaj Dayama, and Vinayaka Pandit, "Zero Knowledge Proofs Towards Verifiable Decentralized AI Pipelines," *IBM Research Lab*, India; Tianyi Liu, Xiang Xie, and Yupeng Zhang, "zkCNN: Zero Knowledge Proofs for Convolutional Neural Network Predictions and Accuracy," *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, November 2021.

Amid the uncertainty as to how complex models reach their decisions—often referred to as the "black box" problem—participants noted the need for an improved understanding of the principles of operation behind these models. This is particularly important in this instance given the dangers of errors in a nuclear context. It is further complicated by the likelihood that any integrated system will have "AI and non-AI parts" that require reconciliation and may have differing risk profiles. Indeed, the risk profiles of AI systems vary qualitatively based on the type of AI model at issue (e.g., computer vision, natural language processing, predictive intelligence), as well as the way in which the model is being utilized. Without proper understanding of the complexities of various AI models as well as the ways in which they would be utilized within nuclear systems, it is neither possible to understand nor make informed decisions about the risks created by integrating them into nuclear architectures. Accurate depiction and fulfillment of these knowledge gaps will help inform the risk-benefit calculus for those considering the adoption of these models.

The significant takeaways focused on the technical concerns surrounding AI technologies included:

» **Development of technical toolkits to arm policymakers** with further information on the intricacies of these technologies. These toolkits could potentially be used to ensure verification and validation of AI-enabled systems in agreements and potentially move confidence building measures (CBMs) forward. They could also include novel ideas such as licenses for harnessing compute resources, hardware-focused CBMs, and zero-knowledge and/or formal methods for verification.

» There remains a significant need for **continued, focused discussion** via trusted venues in which technical experts can engage with national security policymakers in order to flesh out the above-mentioned toolkits. These conversations are currently too rare and must be sustained and expanded to build out a Pugwash-style deep bench of expertise that informs national-level, significantly risky decisions, such as the incorporation of machine learning techniques into NC3.

» **Traditional cyber vulnerabilities**, which establish the broad understanding that everyone should "assume breach," are also inherited by machine learning (ML) systems. This demands a better understanding of the interconnection between these risks and opportunities, with an eye towards next generation systems that will be vastly more reliant on digital components and architectures.

» **AI Red Teams** are too few and far between, and a larger trust community needs to be established for sharing insights, best practices, and injecting more informed understandings of risk and opportunities. This could be realized through information sharing and analysis (ISAC) models for vulnerabilities in AI systems. Similar to cybersecurity red teams, they can be tasked with detecting and uprooting attacks on data and the corresponding ML algorithms. AI Red Teams are an admission that the verification toolkits of the first bullet will always fall short and we will not be able to certify our way to AI Safety. Understanding that gap is important for policy makers and end users of AI in high risk systems.[45]

» The breadth of **potential threats posed by malicious use of ML as well as threats to ML systems themselves** will expand as it becomes clearer how and where ML will be integrated into NC3 and society more broadly. ML attack vectors could be utilized to disrupt countries, target policy decisions through 'sock puppet campaigns,' create a cloud of confusion during a crisis, or target critical personnel, such as servicemembers for nuclear submarines, commanders, politicians, and others engaged in the overall decision making process.

» **Compute power may become equal to political power.** In order to break this cycle, the creation of a governance regime for compute power could aid in ensuring that those with the most resources do not impose their will on those without. This could include an agreement between global powers to control global computing and hardware resources by tracking the development and purchase of compute power—using both licenses for private computing firms and checks on the destinations of physical processors.

Additionally, potential vulnerabilities will be introduced by integrating AI into existing systems. Per our discussions, this could include traditional vulnerabilities associated with cyber attacks (e.g., insider access, overprivileged data and code storage, denial of service), data poisoning, label poisoning, model theft, model evasion, and model inversion. Potential solutions to some of these challenges include:

Ensuring interpretability

The mathematical principles behind machine learning are well understood. But to improve safety, ML must become human interpretable—meaning that operators or users can predict behavior even in unprecedented situations, as safety must be guaranteed.

---

45   This point was contributed by an AI technical participant.

Increasing federal spending on AI safety research

If something goes wrong with a system, it may be difficult, if not impossible, to isolate the error. Despite the significance of the risk posed by these errors, less than 1% of federal AI spending goes to AI safety research. Participants suggested that federal AI safety research should focus on the following four areas: setting standards for AI research and development; creating a test bed for developing AI methods; reducing dependence on single sources of data; and understanding the political drivers associated with including autonomy in second strike systems. This recommendation is made against the backdrop of a talent bottleneck, commercial bottleneck, and semiconductor crunch.

Assessing and improving the security of AI supply chains

The security of AI hardware supply chains is critical, as is determining ways to monitor and sanction compute power. Strategies to help verify compliance include zero-knowledge software review, tracking shipments of computing power (some participants noted the potential for future export control or a safeguards model), supporting a "trusted foundry" model, and creating standards for tools to prove the provenance of data. Participants noted that the "dual-use" nature of these systems and the prevalence of non-state actors driving their development challenge efforts to control the proliferation of AI technologies that might be used in a military context. In addition, it was a prudent point that many of the same cyber vulnerabilities that already exist are being inherited by AI and ML models and therefore continue to be cause for concern.

Creating the conditions for mutual understanding and the sharing of best practices with partners, allies, and—when appropriate—adversaries.

Participants discussed the value of unilateral declarations as well as potential venues for these discussions, including bilateral engagements in the form of strategic stability dialogues, engagement via the P5 process, and Track 1, 1.5, and 2 processes. Participants also discussed the potential role of non-state actors in governing the risks posed by AI-NC3 integration, with some noting the importance of the research community (e.g., Pugwash) in addressing past risks. These discussions would work towards increasing mutual transparency and avoiding "flash wars" akin to the "flash crashes" connected to AI-driven high-frequency trading, where the interactions between AI systems create unintended consequences.

Alongside these technical challenges, we also engaged with CBMs that states might undertake unilaterally, bilaterally, or multilaterally to reduce the stability risks outlined above.

Confidence Building Measures

Throughout the table-top exercise and the group discussions, participants proposed multiple potential CBMs to reduce nuclear strategic stability risks accelerated by advancements in AI technologies.[46] Some of these CBMs are not implicitly about nuclear weapons, but rather nuclear adjacent CBMs that could add to strategic stability.

Proposed CBMs

CBMs that involve agreeing to, or communicating an intent to, renounce or limit the use of AI technologies in certain weapon and military systems, including NC3.

CBMs that encourage governments and industry players to agree on standards, guidelines, and norms related to AI trust and safety, as well as "responsible" use of AI technologies. These could take the form of unilateral statements (or statements by major industry players) or multilateral agreements on definitions and trust and safety approaches, negotiated in either bilateral or multilateral settings.

CBMs that increase lines of communication, such as hotlines and crisis communications links and/or improve the quality, reliability, and security of communications in crisis. This category also includes confidence building measures for crisis response, both in the private and the public sector.

---

46   The proposed CBMs are elaborated on and organized on a sliding scale in Appendix 1: Confidence Building Measures Scales.

CBMs that increase collaboration between private and public industry, as well as between governments, both allied and adversarial. Also included are CBMs that encourage education and training for policymakers, decision makers, and diplomats on AI, with a focus on technical and statistical literacy. Lastly, this area also includes knowledge and best practices sharing in both the public and the private sectors.

Scientists and engineers noted that building robust AI systems is complex, requiring evaluation, assurance, safety, reliability, and interpretability testing processes. Each of these developmental steps can and likely would be short-changed as part of a "race to the bottom" among governments integrating these technologies into their military capabilities. There are also concerns that adversaries might bypass some of the testing and evaluation processes, raising the risks of miscalculation and even accidents associated with AI systems.

To create the assurance needed for decision makers to have confidence in AI systems, participants suggested that states consider adopting globally accepted standards, though the institutional apparatus required to create these standards remains underdeveloped. Greater international convergence on frameworks could increase transparency on testing and evaluation procedures of AI systems. Technical participants emphasized that transparent AI standardization is possible without revealing sensitive information on methods and that globally agreed benchmarking and standardization can increase global reliability and security of AI systems. Furthermore, compliance, adherence, and transparency in the standardization of AI norms and procedures should build trustworthiness in AI systems.

Participants also suggested that those states developing AI technologies for military purposes should agree to take part in dialogues designed to establish shared strategic and cultural understanding of AI risks. As the effects of AI technologies can be unpredictable and can cause disruption, it is important to develop a shared strategic and cultural understanding of the use of algorithmic decision support tools. Of particular concern is the integration of these tools with the NC3 mission. Collaboration and red-teaming among states, industries, academia, and non-governmental bodies are essential to creating a baseline understanding of potential AI-nuclear risks. Participants also noted that the lack of clarity on red-lines related to AI integration in NC3 systems could be highly destabilizing. A focused approach to AI risk reduction lines of effort is necessary to establish strategic stability among nuclear weapons states.

Discussions also saw a focus on domestic responsibility. States could concentrate domestically on CBMs that involve domestic regulations and standards for AI, including supply chains, import-export, standards of use, and rating companies. This area also includes domestic sanctions against those companies or governments that undermine the regulations and standards.

Next Steps

While the research questions pertaining to nuclear stability are expansive, they only skim the surface of the broader strategic challenges facing scientists, engineers, policymakers, and academics as AI technologies progress and become increasingly integrated into military operations.

Of those broader questions, we are particularly interested in the asymmetries in the development, deployment, and use of AI technologies in high stakes military systems– and how this might impact state behavior. The interaction between the technological and contextual factors are also fodder for future work. Indeed, in the TTX, the contextual factors, including geography, alliance relationships, and domestic political imperatives, were considered as important as the technologies at issue for the players. We are also keen to further explore the various types of governance mechanisms available to policymakers at the domestic and international level to address these concerns. In recent months, much has been made of creating an "IAEA for compute," for example. While compute likely represents the most viable AI-related technology for verification, there is much to dislike about the analogy given the reliance of voluntary offer arrangements for a subset of states parties to the IAEA and the fact that a number of countries exist outside of the nuclear nonproliferation regime. The work to address and assess the relative benefits of regulating technology as opposed to specific use cases and to decide whether verification is required in any regime addressing the integration of AI technology in military contexts is just beginning.

We look forward to playing our part in continuing to ask and answer some of these questions and hope that the TTX and brush clearing associated with this project prove useful to the Bureau of Arms Control, Verification, and Compliance moving forward.

Appendix: Confidence Building Measures Scales

The following confidence building measure scales represent the areas in which states may embrace confidence building measures on nuclear weapons and artificial intelligence, as well as a range of variations of those measures based on levels of effort and collaboration. These examples are not an exhaustive list. Rather, this list can grow and improve with additional workshops, further research, and continued interviews with subject matter experts.

Less effort, less international collaboration ←	Confidence Building Measures	→ *More effort, more international collaboration*

AI Use or Non Use in Weaponry and Military Systems

1.1	*Public statements by officials*	**Communication on How Nations Use, or Do Not Use, Artificial Intelligence in NC3 or Delivery Systems**	*Multilateral agreement, communication, or practice*
	Ex: Lt. Gen Shanahan states that the U.S would not have a Dead Hand with artificial intelligence.	**Mid scale options** National declaration Bilateral agreement or commitment Alliance wide practices and standards UN consensus document Conversations between states Track I and Track II dialogues UN working group **Possible communication, agreement, or practice** Limits on artificial intelligence integration into NC3 or certain nuclear/NC3 systems Limits on artificial intelligence in delivery systems Limits on uncrewed nuclear delivery systems Requirements regarding human involvement in launch decisions	Ex: Countries with nuclear weapons sign a multilateral agreement.

1.2	*Public statements by officials*	Communication of, or Agreements on, Not Deploying, or Limits to Deployment, of AI Systems to Contested Areas	*Multilateral agreement, communication, or practice*
	Ex: A defense official states that military systems that use artificial intelligence for decision support will not be deployed to a contested territory.	**Mid scale options** National declaration Bilateral agreement or commitment Alliance wide practices and standards UN consensus document Conversations between states Track I and Track II dialogues UN working group **Possible communication, agreement, or practice** Limits on deploying, or refusal to deploy, autonomous systems in designated contested areas of concern Limits on, or refusal to, deploy advanced AI tools within armed systems	Ex: Nations agree to limit deployments of certain types of artificial intelligence systems during conflict.

AI Trust, Safety, and Responsible Use

2.1	*Public statements by officials*	Communication on Artificial Intelligence Safety and Assurance Practices and Standards	*Multilateral agreement, communication, or practice*
	Ex: A U.S. official states that the U.S. will develop internal TEVV standards for nuclear weapon systems and be bound by the ensuing standards.	**Mid scale options** National declaration Bilateral agreement or commitment Alliance wide practices and standards UN consensus document Conversations between states Track I and Track II dialogues UN working group **Possible communication, agreement, or practice** Agreement on what constitutes TEVV best practices or standards for nuclear systems Sharing of TEVV capabilities or standards with allies and potentially adversaries Conversations on TEVV and AI safety and assurances Acknowledgments that artificial intelligence systems can break and should not be treated as perpetually reliable	Ex: A multilateral agreement between countries sets agreed upon TEVV standards accompanied by verification inspections.

2.2	*Public statements by officials*	**Communication between States on What Constitutes Responsible Artificial Intelligence Use in the Nuclear Domain and Beyond**	*Multilateral agreement, communication, or practice*
	Ex: The CDAO sets out a clear description of how they understand responsible artificial intelligence use.	**Mid scale options** National declaration Bilateral agreement or commitment Alliance wide practices and standards UN consensus document Conversations between states Track I and Track II dialogues UN working group **Possible communication, agreement, or practice** Responsible artificial intelligence best practices Standards on responsible artificial intelligence Explanation of how a country understands responsible artificial intelligence use	Ex: A signed multilateral agreement establishes a set of principles for responsible intelligence use.
2.3	*Public statements by officials*	**Human - AI System Interaction Norms**	*Multilateral agreement, communication, or practice*
	Ex: An official states that all nuclear launch related AI systems will have a human in the loop.	**Mid scale options** National declaration Bilateral agreement or commitment Alliance wide practices and standards UN consensus document Conversations between states Track I and Track II dialogues UN working group **Possible communication, agreement, or practice** Human in the loop for NC3 systems Human briefing of some outputs from artificial intelligence systems Practices or standards for human-machine hybrid analysis and intelligence	Ex: A multilateral agreement requires there to be a human in the loop for certain NC3 systems.

2.4	*Public statements by officials*	**Security Communications and Agreements**	*Multilateral agreement, communication, or practice*
	Ex: An official states that their department will collaborate with the private sector to better detect and respond to data poisoning attacks.	**Mid scale options** National declaration Bilateral agreement or commitment Alliance wide practices and standards UN consensus document Conversations between states Track I and Track II dialogues UN working group **Possible communication, agreement, or practice** Norms on avoiding the data poisoning of safety critical data streams Agreed upon standards to combat data poisoning Standardization and adversarial testing Cooperation on data poisoning detection Private-public security partnerships and information sharing	Ex: A multilateral agreement is reached on security standards to combat data poisoning.

Improved Communication and Crisis Response

3.1	*Informal communication by officials*	**Effective Communication Channels**	*Multilateral communication mechanisms*
	Ex: An American ambassador sends a WhatsApp message to their counterpart about an incident related to an AI weapons system.	**Mid scale options** Theater level communication channels Operational communication channels Head of State communication channels **Possible communication, agreement, or practice** Hotlines Joint data centers Risk reduction centers Doctrine exchanges	Ex: The U.S. president uses a multilateral hotline to speak to his Russian and Chinese counterparts about an incident related to an AI weapons system.

3.2	*Department or agency crisis response teams*	**Artificial Intelligence Emergency Readiness Teams**	*Multilateral crisis response teams*
	Ex: An agency creates a crisis response team for geopolitical crises and incidents related to artificial intelligence.	**Mid scale options** Expansion of national CERTs to include AI incidents National crisis response team Bilateral crisis response team Alliance-wide crisis response team Public-private joint crisis response team	Ex: Several nations agree to stand up crisis response teams and permit them to interact during times of crisis or several nations agree to leverage national CERTs and the cross-border relationships they already have, expanding them to include AI incidents.

3.3	*Public statements by officials*	**Attribution and Adjudication for Cyber Events Affecting Artificial Intelligence**	*Multilateral attribution and adjudication center*
	Ex: An official attributes an attack on an artificial intelligence system to an adversary.	**Mid scale options** National attribution center Bilateral attribution center Alliance-wide attribution center Public-private joint attribution center	Ex: A government stands up an international digital forensics or technical investigatory body to attribute cyberattacks on artificial intelligence.

AI Collaboration, Education, Transparency, and Knowledge Sharing

4.1	*Informal discussions between private sector individuals*	**Private Sector Communication and Collaboration**	*Formal forums for international private sector collaboration*
	Ex: Private sector researchers from two allied countries informally interact and share information.	**Mid scale options** Working groups within bilateral relationships Working groups within alliances Track II dialogues Formal private-public collaboration **Possible communication or agreement** Conversations between private industry scientists and engineers and government officials Responsible AI collaboration between industry and government Shared research norms with adversary researchers and academics	Ex: The UN establishes a formal working group of private sector representatives dedicated to AI collaboration, education, transparency, and knowledge sharing.

4.2	*Department or agency education programs*	AI Education for Policy and Decision Makers	*Multilateral working groups and knowledge exchanges*
	Ex: A department regularly organizes educational discussions and briefs from AI experts for their employees.	**Mid scale options** Visible attempts by politicians and policymakers to understand the more difficult areas of artificial intelligence, demonstrating that leaders are aware that AI is not just magic **Possible communication, agreement, or practice** Education on existing and future integrations of AI into nuclear weapon systems Education on adversary and allied AI progress and developments Education on AI safety and alignment Education on AI vulnerabilities and attack vectors	Ex: A multilateral group of officials and AI researchers regularly meets to discuss new AI developments, threats, and safety research.
4.3	*Public statements by officials*	Doctrine and Process Sharing	*Multilateral agreement, communication, or practice*
	Ex: An official publicly defines strategic stability and how they view artificial intelligence impacting strategic stability.	**Mid scale options** National declaration Bilateral agreement or commitment UN consensus document Conversations between states Track I and Track II dialogues UN working group **Possible communication, agreement, or practice** Doctrine exchanges or agreement on interpretations of strategic stability Doctrine exchanges or agreement on artificial intelligence definitions and understanding Shared definitions of artificial intelligence applications	Ex: Countries exchange multilateral communication on shared definitions of artificial intelligence applications and threats.

INSTITUTE FOR SECURITY AND TECHNOLOGY
www.securityandtechnology.org

info@securityandtechnology.org