Avin, Shahar and S.M. Amadae, "Autonomy and Machine Learning as Risk Factors at the Interface of Nuclear Weapons, Computers and People," *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Euro-Atlantic Perspectives*, SIPRI, May 2019, 105-118.

https://www.sipri.org/publications/2019/other-publications/impact-artificial-intelligence-strategic-stability-and-nuclear-risk

Avin, Shahar and S.M. Amadae, “Autonomy and Machine Learning as Risk Factors at the Interface of Nuclear Weapons, Computers and People,” The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Euro-Atlantic Perspectives, ed. by Vincent Boulanin, SIPRI, May 2019, 105-118.

12. Autonomy and machine learning at the interface of nuclear weapons, computers and people

SHAHAR AVIN AND S.M. AMADAE[1]

A new era for our species started in 1945: with the terrifying demonstration of the power of the atom bomb in Hiroshima and Nagasaki, Japan, the potential global catastrophic consequences of human technology could no longer be ignored. Within the field of global catastrophic and existential risk, nuclear war is one of the more iconic scenarios, although significant uncertainties remain about its likelihood and potential destructive magnitude.[2] The risk posed to humanity from nuclear weapons is not static. In tandem with geopolitical and cultural changes, technological innovations could have a significant impact on how the risk of the use of nuclear weapons changes over time.

Increasing attention has been given in the literature to the impact of digital technologies, and in particular autonomy and machine learning, on nuclear risk. Most of this attention has focused on ‘first-order’ effects: the introduction of technologies into nuclear command-and-control and weapon-delivery systems.[3] This essay focuses instead on higher-order effects: those that stem from the introduction of such technologies into more peripheral systems, with a more indirect (but no less real) effect on nuclear risk. It first describes and categorizes the new threats introduced by these technologies (in section I). It then considers policy responses to address these new threats (section II).

I. New technology brings new threats

The risks of these higher-order effects can be divided into two categories.

1. In the first category are new vulnerabilities in the trusted computing base (TCB) of nuclear deterrence due to the introduction of machine learning into

[1] The authors would like to thank the participants in the Plutonium, Silicon and Carbon Workshop held by the University of Cambridge Centre for the Study of Existential Risk in Sep. 2018, for a lively discussion of these topics. They are also grateful to Jon Lindsay for sharing unpublished materials and insights, and to Vincent Boulanin, Baruch Malewich and Liran Renert for helpful comments.

[2] For an estimate see e.g. Barrett, A. M., Baum, S. D. and Hostetler, K., ‘Analyzing and reducing the risks of inadvertent nuclear war between the United States and Russia’, *Science & Global Security*, vol. 21, no. 2 (2013), pp. 106–33, <https://doi.org/10.1080/08929882.2013.798984>, p. 120.

[3] Thompson, N., ‘Inside the apocalyptic Soviet doomsday machine’, *Wired*, 21 Sep. 2009, <https://www.wired.com/2009/09/mf-deadhand>; and ‘“Doomsday machine”: Russia’s new weapon reportedly gets nuclear warhead’, Sputnik, 17 May 2018, <https://sputniknews.com/russia/201805171064549993-russiaposeidon-system-torpedo/>. See also the other chapters, in particular 4–10 and 13, in this volume.

nuclear command, control, communications, computers, intelligence, surveillance and reconnaissance (NC4ISR) systems. The TCB of a computer system is 'The totality of protection mechanisms within [that] system . . . responsible for enforcing a security policy'.[4] Nuclear deterrence presumably requires a security policy that always allows authorized personnel to (*a*) detect threats that call for a nuclear response and (*b*) launch a nuclear response, while (*c*) never allowing unauthorized personnel to launch nuclear weapons. As such, at a minimum, the TCB of nuclear deterrence would include all critical NC4ISR systems (i.e. those systems where a malfunction or compromise would undermine *a*, *b* and *c*).

2. The second category of risks consists of novel and amplified threats from the use of autonomy and machine learning in the planning and execution of cyber operations and influence campaigns against nuclear weapon systems and associated personnel.

Both of these categories expand and amplify existing threats, rather than introduce entirely new categories of threat. Nonetheless, the scale of the effect is substantial and may render feasible certain attacks that were previously infeasible.

Machine learning and autonomy in NC4ISR introduces new attack surfaces

Computer systems are susceptible to attack. They rely on many lines of code that contain numerous opportunities for developers to make a mistake or fail to consider all possible implications, in a way that introduces a vulnerability—a bug. A patient and resourceful adversary is often able to reliably find and exploit such vulnerabilities in order to gain control of or disrupt the operations of a computer or computer-based system.

Responses to this computer security threat have evolved over the decades, from pre-deployment testing to formal guarantees that certain parts of code do not contain specific kinds of vulnerability.[5] Another powerful practice is to limit the 'attack surface' of a system—that is, all the points at which an attacker can interact with the systems. For instance, this can be done by restricting functionality, introducing authority restrictions or restricting input channels, or through practices such as air-gapping, which physically separates the system from any network.[6] However, some of these security practices limit autonomy, which requires a high-level of functionality and integration with numerous inputs (including networked resources). Thus, wherever there is a

[4] US Department of Defense (DOD), *Department of Defense Trusted Computer Systems Evaluation Criteria*, DOD Standard 5200.28-STD (DOD: Washington, DC, 26 Dec. 1985), <https://apps.dtic.mil/dtic/tr/fulltext/u2/a207905.pdf>, p. 116.

[5] Anderson, R., *Security Engineering: A Guide to Building Dependable Distributed Systems, Second Edition,* (Wiley Publishing, Inc.: Indianapolis, IN, 2008), chapter 26.

[6] Saltzer, J. H. and Schroeder, M. D., 'The protection of information in computer systems', *Proceedings of the IEEE*, vol. 63, no. 9 (Sep. 1975), pp. 1278–1308, <https://doi.org/10.1109/PROC.1975.9939>.

push towards autonomy that allows for complex behaviour (e.g. human-like or even animal-like perception or behaviour), these security practices may not be viable.

The challenge of maintaining computer security against digital attacks is even harder for machine learning than for autonomy. While autonomous complex behaviour could be produced through a set of rules laid out and scrutinized by a developer, a machine learning approach to a problem instead seeks to bring about correct behaviour through analysis of large amounts of data. While the learning algorithm is specified, scrutinized and tested by the developer, the learned behaviours in many contemporary approaches cannot be scrutinized to the same degree as rule-based systems.[7]

It is already known that a broad range of models trained through machine learning are susceptible to a new kind of vulnerability, termed 'adversarial examples': an adversary can craft a malicious input that reliably causes a trained model to produce the wrong behaviour (e.g. misclassify an object in an image or take an inappropriate action in the environment).[8] While this vulnerability has been known and researched heavily for several years, no robust solution has yet been found. Nonetheless, given the promise of new capabilities that machine learning and automation offer, the pressure to deploy potentially insecure systems may present itself.[9]

When considering threats that might be introduced from increased autonomy and use of machine learning, it is important to consider the entire sprawling range of systems and functions that make up and support NC4ISR. Specific attention has been given to delivery systems and to nuclear command, control and communications (NC3).[10] The awareness of potential threats to 'core' computer systems in NC3 has led to significantly improved security for such systems, and some reluctance to introduce autonomy and machine learning into them.[11] However, more peripheral systems can also pose a threat, especially as they are more likely sites for the introduction of autonomy and machine learning. These include, for example, systems onboard satellites that relay communications and images or the simulators used to plan and test strategies. They can also extend as far as the vast computer systems and

[7] Barreno, M. et al., 'The security of machine learning. *Machine Learning*, vol. 81, no. 2 (Nov. 2010), 121–48, <https://doi.org/10.1007/s10994-010-5188-5>.

[8] Szegedy, C. et al., 'Intriguing properties of neural networks', arXiv, 1312.6199, version 4, 19 Feb. 2014, <https://arxiv.org/pdf/1312.6199v4.pdf>.

[9] Geist, E. and Lohn, A. J., *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (Rand Corporation: Santa Monica, CA, 2018), <https://www.rand.org/content/dam/rand/pubs/perspectives/PE200/PE296/RAND_PE296.pdf>, p. 10.

[10] On delivery systems see US Government Accountability Office (GAO), *Weapon Systems Cybersecurity: DOD Just Beginning to Grapple with Scale of Vulnerabilities*, GAO-19-128 (GAO: Washington, DC, 9 Oct. 2018), <https://www.gao.gov/assets/700/694913.pdf>. On NC3 see Anderson, R., *Security Engineering: A Guide to Building Dependable Distributed Systems, Second Edition,* (Wiley Publishing, Inc.: Indianapolis, IN, 2008), chapter 13. On these issues see also chapters 5–10 and 13 of this volume.

[11] On the vulnerability of machine learning to cyberattack as an obstacle to the adoption of machine learning in the military sphere see also chapters 3 and 6 of this volume.

networks that provide news information to the public and to civilian officials, which may affect tactical or strategic decision-making.

Admittedly, it is not always easy to chart a scenario that begins with a compromise of a particular peripheral system and ends with the unauthorized launch of a nuclear weapon.[12] It is similarly difficult to describe a scenario whereby an adversary would intervene in the authorized launch of a nuclear weapon. However, these systems are present for the well-funded and patient adversary to explore and exploit. In particular, there is increasing concern about attacks that initially target command, control, communications, computers, intelligence, surveillance and reconnaissance (C4ISR) systems that are 'entangled'—that is, used for both nuclear and conventional weapons—such as satellites, intelligence gathering and logistics.[13] Entangled systems present two challenges here: first, they are often not considered 'nuclear' systems, so are subject to a lower level of security scrutiny than nuclear systems. Second, attacks on such systems may be considered by an adversary as unlikely to trigger a nuclear escalation, leading to a miscalculation—the adversary may not even know that the system has an NC4ISR purpose, and therefore consider the attack to be conventional, while the targeted state may perceive the attack as an attack on its nuclear capabilities.

Machine learning and autonomy can be used to carry out cyber and influence operations against nuclear systems and personnel

Having surveyed ways in which a state may heighten vulnerability and risk by introducing autonomy and machine learning into its own NC4ISR systems, the various ways in which an attacker could deploy machine learning and autonomy to compromise an adversary's NC4ISR systems—even those that do not feature any autonomy or machine learning—are now considered.

The attack surface of the NC4ISR systems of a nuclear-armed state is composed of numerous computer systems (as surveyed above) and also a broad range of personnel. These include the military personnel in charge of deploying weapons; the civilian contractors tasked with building and maintaining weapon systems; and the civilian authorities that take decisions to fund maintenance, modernization or retirement of weapon systems. There are also the individuals, groups and international bodies that advocate arms control measures and seek to sway public opinion and nuclear norms, and many others on the long list of involved persons.

No computer system should be considered perfectly secure. Rather, security mechanisms are placed to increase the cost or the risk to the attacker to a level that makes an attack effectively impractical under most expected conditions. For example, requiring the simultaneous action of two individuals to arm a

[12] For an in-depth exploration of this see Futter, A., *Hacking the Bomb: Cyber Threats and Nuclear Weapons* (Georgetown University Press: Washington, DC, 2018).

[13] Acton, J. M., 'Escalation through entanglement: how the vulnerability of command-and-control systems raises the risks of an inadvertent nuclear war', *International Security*, vol. 43, no. 1 (summer 2018), pp. 56–99, <https://doi.org/10.1162/isec_a_00320>.

nuclear weapon requires an attacker to compromise two insiders instead of one. Air gapping a system requires an attacker to gain physical access to the system. Within narrow domains, cryptography and computer security can ensure that the computational power required to attack a system is astronomical. However, when considering the entire attack surface of NC4ISR, it is not currently possible to provide such guarantees for the system as a whole. In theory, applications of machine learning and autonomy on the attacker's side can reduce the cost of an attack and transform the target system system from being 'effectively secure' to being 'effectively insecure'.

Articulating the specific ways in which autonomy and machine learning could reduce the cost of an attack requires access to information that is partly or entirely classified. Instead, the kinds of novel attack that nuclear-armed states should consider in their threat assessments are illustrated by the following two qualitative descriptions of scenarios that feature autonomy and machine learning in numerous places within an attacker's system.[14]

Use of machine learning and autonomy to compromise NC4ISR computer systems at scale

In this scenario, country A is interested in developing a reliable capability to monitor, degrade or disrupt numerous key digital components of country B's NC4ISR systems. First, country A finds information about potential targets in country B's systems, for example, what hardware and software are installed, the network setup and access, and so on. This is traditional intelligence work: gathering information from sources in procurement, defence contractors and in military bases.[15] Country A may deploy machine learning to process large volumes of mostly irrelevant data from commercial, trade, procurement, budgetary or logistics sources that may shed light on which systems are installed and where. If country A is well positioned to do so, it may aim to become the upstream supplier of components for its adversaries' military systems.[16]

Once a list of target technologies is compiled, country A can gain access to country B's systems via a copy of either compiled or source code or through a remote connection or a replica. With access to source code, country A can search for vulnerabilities in the target systems and create exploits.[17] Machine learning techniques and automation expedite the search for patterns of common mistakes that could lead to an exploit. Access to compiled code,

[14] On the potential use of machine learning in attacks see Brundage, M. et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation* (Future of Humanity Institute et al.: Oxford, Feb. 2018), <https://arxiv.org/pdf/1802.07228v1.pdf>.

[15] Sanger, D. E., *The Perfect Weapon: War, Sabotage, and Fear in the Cyber Age* (Crown: New York, 2018).

[16] E.g. Robertson, J. and Riley, M., 'The big hack: how China used a tiny chip to infiltrate U.S. companies', Bloomberg Businessweek, 4 Oct. 2018, <https://www.bloomberg.com/news/features/2018-10-04/the-big-hack-how-china-used-a-tiny-chip-to-infiltrate-america-s-top-companies>.

[17] Jon B. and Rich T., 'A day in the life of an NCSC vulnerability researcher', British National Cyber Security Centre, 17 Nov. 2017, <https://www.ncsc.gov.uk/blog-post/day-life-ncsc-vulnerability-researcher>.

when combined with reverse engineering, allows a similar machine learning- and automation-expedited search for vulnerabilities. Finally, with only 'black box' access to a system (where inputs can be sent to the system and outputs can be read out, but no access to source or compiled code is possible), security researchers can try a large number of input combinations to find vulnerabilities. This method, called 'fuzzing', is often heavily automated.[18]

Once country A has identified a range of vulnerabilities in country B's systems, it needs to devise a plan for how to use them. To make detection harder and increase deniability, country A might take control of a third party's insecure computational resources to set up autonomous or semi-autonomous bots armed with the code needed to launch the exploits against country B's systems. The systems that control the network of bots may themselves include significant automation, to allow many computers to operate in synchronization and to further complicate detection and attribution. Machine learning tools could be used to analyse the statistical profile of traffic in the target network or intermediary networks, so that bot-generated traffic could mimic the same distribution and avoid detection by statistics-based defence tools.

In these examples, autonomy and machine learning do not present country A with an entirely novel capability, but instead increase the scale of existing capabilities or reduce the costs of staff and training. In addition, autonomy and machine learning increase distance and reduce the likelihood of discovery and attribution. In doing so, they may lower the perceived costs (in money or fear of retaliation) of an attack.

Use of machine learning and autonomy to launch a nuclear-capability-retarding manipulation campaign

In this scenario, country A seeks to influence opinions and decision-making in country B. This might be by decreasing the funds and talent available for country B's nuclear operations or by decreasing the likelihood that country B will respond to an ambiguous or threatening situation with a nuclear attack.

First, country A identifies the decision makers it would like to influence. These could be elected officials who vote on budgets, senior military personnel who decide on future plans and protocols for escalation, or potential recruits who decide on whether to pursue a career in the nuclear apparatus. Next, country A maps the opinions and beliefs that guide individuals' decisions, maps the sources through which these opinions and beliefs are shaped, and determines which will be possible for an outsider to shift. Opportunities for influence often present themselves when large and technology-engaged publics are involved or when free and open discussion is valued.[19]

[18] Sutton, M., Greene, A. and Amini, P., *Fuzzing: Brute Force Vulnerability Discovery*, (Pearson Education, Inc.: Upper Saddle River, NJ, 2007).

[19] Lin, H. and Kerr, J., 'On cyber-enabled information/influence warfare and manipulation', 8 Aug. 2017, to appear in *Oxford Handbook of Cybersecurity* (Oxford University Press: Oxford, forthcoming), <https://ssrn.com/abstract=3015680>.

At this point, country A can profile its targets and identify the intermediary influencers it would need to engage.[20] To profile a target, it might study the target's behaviour (e.g. websites that she or he visits) and her or his identity and group membership, other beliefs and ideologies (from public statements), then draw up a psychological profile, and so on. Such information can be accessed today by several private companies (e.g. Facebook, Twitter) and the advertising firms that work with them. Based on the established profiles, country A can begin an influence campaign with a trial-and-error method of testing and refining targeted content (e.g. adverts, direct messages, news stories, etc), all the while measuring engagement and the magnitude of the effect that the messages have on the targets' behaviour. This method significantly benefits from automation, and particularly from the ability to tailor messages that drive each individual towards the desired behaviour. The paths to that behaviour may be different for each person.[21] For example, risk-averse individuals or communities might be targeted with historical evidence of nuclear accidents, while small-government oriented communities might be advised of the costs of maintaining nuclear deterrence. Prospective recruits could be targeted with alternative job offers or careers.

A nascent and powerful influencing technology is the ability to create life-like forgeries of faces using generative adversarial networks (GANs). This enables the creation of videos in which individuals appear to be saying things that they have not said.[22] These may be particularly powerful in reinforcing ideas to which a target community is ideologically predisposed. Forensic methods to identify content as fake are in their infancy and their efficacy is still in doubt.[23]

The two threat scenarios outlined above—a search for vulnerabilities in an adversary's nuclear digital information systems and influence campaigns to alter an adversary's nuclear readiness and resolve—have existed since the Cold War era. However, both contain numerous steps that can be facilitated by autonomy and machine learning. These may lower the cost to the attacker, increase the speed, scale and efficacy of an attack, or reduce the risk to the attacker by obfuscating the links to the source and allowing for plausible deniability. This aspect of machine learning and autonomy in the nuclear weapons domain should be explored and red-teamed by parties who are in a

[20] Kosinski, M. et al., 'Mining big data to extract patterns and predict real-life outcomes', *Psychological Methods*, vol. 21, no. 4 (Dec. 2016), pp. 493–506, <https://doi.org/10.1037/met0000105>.

[21] Cai, H. et al., 'Real-time bidding by reinforcement learning in display advertising', *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Association for Computing Machinery (ACM): New York, 2017), pp. 661–70, <https://doi.org/10.1145/3018661.3018702>.

[22] Suwajanakorn, S., Seitz, S. M. and Kemelmacher-Shlizerman, I., 'Synthesizing Obama: learning lip sync from audio', *ACM Transactions on Graphics*, vol. 36, no. 4 (2017), article no. 95, <https://doi.org/10.1145/3072959.3073640>. On GANs see also chapter 1 in this volume. On the malicious use of deepfakes see also chapter 9 of this volume.

[23] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M., 'Faceforensics: A large-scale video dataset for forgery detection in human faces.' arXiv, 1803.09179, 24 Mar. 2018, < https://arxiv.org/pdf/1803.09179.pdf>.

position to access the relevant classified information. Other general policy responses that may be appropriate are considered below.

II. New threats require new policy responses

Cyber threats undermine nuclear deterrence

Nuclear deterrence works to counter threats of either nuclear or large-scale conventional attack through the transparency of its posture. It relies on an always/never alert status: always ready to be executed via legitimate authority, and never subject to compromise. Cyber operations, like other covert actions, rely on stealth: digital attacks exploit vulnerabilities unknown to the target states and often aim to remain secret; once made evident, attackers can lose their advantage as the target state can take counteraction.

Increasingly, cyber vulnerabilities challenge nuclear deterrence because nuclear-armed states may not know that their capabilities have been impaired, and additionally they may have uncertainty about the status of their NC3 or NC4ISR systems. This can lead to either a false sense of confidence and recklessness in issuing threats with escalatory potential, or it may contribute to an overblown sense of vulnerability that encourages pre-emptive action. The erosion of credibility due to these uncertainties also undermines the overarching aim of deterring conflict. Furthermore, the bar to achieving offensive cyber capabilities is much lower than the bar required to establish credible nuclear deterrence, in terms of resources, talent and international regulation. In a world where cyber capabilities can be seen as offsetting nuclear capabilities, the number of potentially relevant actors grows significantly, and the adequacy of existing dyadic nuclear deterrence relations is thrown into question. Thus, the new reality of cyber vulnerabilities, and the tempting advantages to be gained through cyber operations, have created an unprecedented development in warfare.

Recent reports by the US Government Accountability Office (GAO) and the Nuclear Threat Initiative (NTI) have found that even US military systems, which include networked components necessary for operating nuclear armed forces, are vulnerable to electronically mediated digital attacks.[24] Even if, at least within the technologically advanced nuclear-armed states, NC3 systems are fully robust against digital intrusion, it will be more challenging to maintain this impervious systemic integrity with upgrades beyond the original analogue configurations to new digital platforms. Moreover, as itemized by the vast list of potential exploits provided by the GAO's assessment, the entanglement of the nuclear and conventional planning and execution systems suggests that, in order to maintain a credible alert posture, the peripheral

[24] US Government Accountability Office (note 10), p. 30; and Stoutland, P. O. and Pitts-Kiefer, S., *Nuclear Weapons in the New Cyber Age*, Report of the Cyber-Nuclear Weapons Study Group (Nuclear Threat Initiative: Washington, DC, Sep. 2018), <https://www.nti.org/media/documents/Cyber_report_finalsmall.pdf>.

Shahar Avin 21
Deleted: 345

intelligence, surveillance and reconnaissance (ISR) architectures must be maintained at high levels of reliability.[25]

Taken across the board, risks are posed by upgrades to NC3, entangled conventional and nuclear C4ISR systems, budget constraints, difficulties recruiting personnel with appropriate skills, inefficiencies inherent in large bureaucracies with competing jurisdictions, restrictions on sharing information across agencies, and the ongoing advancement of complexity consistent with rapid technological progress. There is an ongoing effort to revamp nuclear weapon systems to be consistent with state-of-the-art technologies, which now include autonomy and machine learning.

The above combination of threats to nuclear deterrence, including the heightened cyberthreats from autonomy and machine learning, call for policy responses.

Deterrence is likely to be insufficient as a policy response

Within the framework of strategic stability, it may seem reasonable to tackle a novel threat, in this case that of the cyber compromise of NC4ISR systems, with deterrence. This is suggested, for example, in the 2018 US Nuclear Posture Review, which presents the threat of nuclear retaliation as a deterrent against cyberattack.[26] However, the wisdom of this approach is questionable.

Historically, deterrence has not proven effective against intelligence collection, special operations and similar covert actions, which cyber operations resemble. Furthermore, adding another trigger for nuclear response and escalation creates one more pathway to catastrophic outcomes through miscalculation or false alarms. For cyberthreats, there is significant uncertainty about the ability of a defender to detect an attack, identify it as an attack and attribute it correctly.[27] Thus, to tackle the threats discussed above, only policy responses other than new forms of deterrence are considered.

Proposed unilateral policy responses

The first class of policy responses involve actions that a nuclear-armed state can take unilaterally to reduce the risks that it is exposed to from digital threats. By making itself more secure, such a state also helps maintain the deterrence relationships that it has in place. Overall, knowledge of potential exploits and steps to avoid them, detect them and address them must be in place as with any other standard security protocols.

According to GAO reports, the US Department of Defense is only beginning to realize the extent of its cyber vulnerability challenges, and the GAO does

[25] Acton (note 13).

[26] US Department of Defense (DOD), *Nuclear Posture Review* (DOD: Washington, DC, Feb. 2018), <https://media.defense.gov/2018/Feb/02/2001872886/-1/-1/1/2018-NUCLEAR-POSTURE-REVIEW-FINAL-REPORT.PDF>.

[27] Lindsay, J. R., 'Restrained by design: the political economy of cybersecurity', *Digital Policy, Regulation and Governance*, vol. 19, no. 6 (2017), pp. 493–514, <https://doi.org/10.1108/DPRG-05-2017-0023>.

Shahar Avin 21
Deleted: 348

not offer any recommendations.[28] Public information about the state of cyber vulnerabilities in other nuclear-armed states is currently lacking. However, it is possible to identify measures to address vulnerabilities from the domain of digital information, computation and communications technologies more generally.

The unilateral policy proposals are surveyed in table 12.1. There are four key points.

1. The integration of information and communications technology (ICT) systems into NC4ISR should be restricted. In particular, the introduction of autonomy and machine learning into these systems should be avoided. This should be reflected in procurement policies.
2. The nuclear-armed states should be mindful of the potential threats and take proactive action to harden systems, enforce security protocols, regularly exercise and simulate attack.
3. These states should develop attribution capacity, adopt procedures and doctrines that increase response time, and plan for rapid recovery from attacks.
4. Good practice should be codified and widely dispersed to relevant personnel. Contingency protocols should be set up, tested and enforced.

The recommendation against introducing autonomy and machine learning into NC4ISR systems should be highlighted, with emphasis heightened in relation to the closeness of a component to critical decision-making or to command and control. The proposals here endorse the NTI report recommendation against integrating these digital capacities into the technical infrastructure necessary to run nuclear security programmes.[29] There will probably be efforts to introduce autonomy and machine learning into conventional ISR. However, due to entanglement, it is a sensible precaution to either severely restrict these methods or, at a minimum, to perform cost–benefit analysis and comprehensive risk assessment. These precautions would allow informed decisions to be made that minimize the erosion of deterrence credibility and the resulting additional risk of inadvertent use of nuclear weapons.

Table 12.1. Unilateral policy responses to reduce nuclear risks from cyber threats

Target	Protect (defence in peacetime)	Detect (response to probing)	Respond (response to attack)
Decision procedure[a]	Strict protocols Secure communication Increase decision time No autonomy; no ML	Routine tests, simulations Attribution capacity	Quarantine Attribute Evaluate Neutralize Counter

[28] US Government Accountability Office (note 10), preface.
[29] Stoutland and Pitts-Kiefer (note 24), p. 8.

Shahar Avin 21
Deleted: 360

			Upgrade
NC3	Redundancy Expertise Secure sourcing System isolation Enhance survivability Cyber resilience Formally verified Cryptographic guarantees Acquisition guidelines No autonomy; no ML No external contracts	Monitoring Testing Attribution capacity	Quarantine Use backup Protocol Buy time Attribute
Nuclear ISR	Redundant sensors Diverse phenomena Intelligence fusion No autonomy; no ML No external contracts	As above	Quarantine Decouple Use backup Protocol Attribute Evaluate Neutralize Counter Upgrade
Conventional ISR	Cost–benefit analysis Comprehensive risk assessment If entangled, treat as nuclear	Monitor (can use ML) Long-term testing Attribution capacity	As above
Nuclear/military personnel	Competitive career opportunities Vet Training, risk awareness Protocols to protect at work and at home Assist in maintaining security	Confidence building Routine checks Monitor (can use ML) Practice attacks Attribution capacity	Counter Attribute Expose Restrict
Public opinion	Education Establish trust Inform about risks Collaborate with media	Monitoring (can use ML) Attribution capacity Counterintelligence	As above
Infrastructure	Upgrade cyber-defences Use cost–benefit analysis, risk assessment to prioritize high-value assets	Enhance industry standards Monitoring (can use ML) Attribution capacity	Assess Bypass Rebuild Upgrade
Supply chain	Trusted sources Security requirements Enhance industry best practice Bespoke systems for NC3	Monitoring (can use ML) Testing Attribution capacity	Assess Report Recall Upgrade
Research and development, testing, simulation,	Procurement guidelines Budget for security Expert, vetted and valued	Oversight, accountability Counterintelligence	Evaluate Counter Attribute

maintenance	staff
	Redundancy

ISR = intelligence, surveillance and reconnaissance; ML = machine learning; NC3 = nuclear command, control and communications.

[a] These include e.g. crisis intelligence and assessment and nuclear planning systems.

Proposed coordination-based policy responses

Coordination around non-use of cyber capabilities against nuclear systems and personnel

A responsible nuclear-armed state can realize that introducing autonomy and machine learning would probably increase its own vulnerability to mistakes and cyber operations, and therefore can unilaterally avoid introducing such methods into NC3 and NC4ISR systems. However it cannot unilaterally prevent another actor from using autonomy and machine learning as tools that enhance cyber operations and influence campaigns. The two scenarios presented in section I demonstrate how autonomy and machine learning could be potentially useful tools in operations against digital systems on the periphery and against individuals and communities of civilians (especially in densely digitally networked societies). It is evident that cybersecurity poses a grave challenge for a country's own nuclear deterrent credibility. It is additionally clear that engaging in offensive operations against other states will necessarily erode the credibility of their nuclear deterrence. To maintain nuclear deterrence and strategic stability, states should exercise restraint by refraining from cyber operations against all other states' nuclear weapons systems and personnel (i.e. broadly targeted information campaigns that could influence nuclear deterrence) and should strive to establish norms and institutions that prohibit such actions.

Costly vigilance and recognition that computerized systems cannot be 100 per cent secure is not unique to either nuclear or conventional military security. This a problem faced across the board in the densely networked digital systems that run finance and banking, communications, air and marine traffic, and healthcare organizations.[30] However, the level of destructive capacity and existential risk is highest for NC3 and NC4ISR systems. Even though attacks in the nuclear domain may be much costlier for the attacker to execute than in other domains, they are not beyond the resources available to states, potentially including non-nuclear-armed states with advanced technology.[31] Assuming that the chief aim of nuclear-armed states in maintaining nuclear deterrence is stability and security—which is contradicted by nuclear war with an inherent perceptible risk of escalation—then no matter

[30] US Government Accountability Office (note 10), p. 30; and Lindsay, J. R., 'Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack', *Journal of Cybersecurity*, vol. 11, no. 1 (Sep. 2015), pp. 53–67, <https://doi.org/10.1093/cybsec/tyv003>.

[31] Slayton, R., 'What is the cyber offense–defense balance? Conceptions, causes, and assessment', *International Security*, vol. 41, no. 3 (winter 2016/2017), pp. 72–109.

Shahar Avin 21
Deleted: 345

how tempting it may seem to disrupt a nuclear-armed state's nuclear command-and-control and related systems, such action risks everyone's security.

Given the common interest in maintaining stability and avoiding nuclear war, all states and global populations stand to gain from constraints against initiating offensive campaigns against military information systems and personnel. Even within civilian societies that are networked and globalized using digital platforms that recognize no national borders, there is a collective benefit to maintaining the cyber commons that rely on cooperation and the development of norms against cyberattacks. This holds even more powerfully when considering the potential weaponization of autonomy and machine learning as a force multiplier for cyber operations and influence campaigns, and the need for norms to prevent such weaponization. Nuclear weapons states share a common interest in developing three basic norms: (*a*) to achieve best practices in maintaining the security of their own NC3 and NC4ISR systems, (*b*) to denounce and refrain from conducting offensive cyber actions, and (*c*) to limit the weaponization of autonomy and machine learning, especially in the cyber and information warfare domains. These limitations should extend beyond the immediate nuclear or military domain, as techniques and methods can be easily transferred from one domain to another.

It seems obvious to develop these cooperative norms among allies, at least those around non-use, because alliance and collaboration is contradicted by either detecting others' cyber vulnerabilities without sharing that information or with the intent to possibly exploit those weaknesses.

However, overall the added risk posed by undermining nuclear deterrence threatens the security of all. A plausible case thus exists to coordinate efforts to prevent development of offensive cyber capabilities (especially highly effective tools that rely on autonomy and machine learning), not only with allied states but also potentially with those whose interests are only partially aligned at best. For example, although China, Russia and the USA do not have the same geopolitical or economic interests, none would benefit from a nuclear conflict.

Coordination around enhanced security and best practices in NC4ISR

In addition to coordination around minimizing offensive use of cyber capabilities (including ones based on autonomy and machine learning), it might also be possible and necessary for nuclear-armed states to coordinate on increased cyber-defences, through the sharing of information about best practices and related defensive technologies. Even despite the impossibility of achieving 100 per cent security in the contemporary world of advanced computation, significant improvements can be made to increase the cost for a putative attacker, at times (e.g. through cryptographic means) to levels that render certain attacks infeasible in practice.

Given that it is, for example, the USA that could lose the most if some aspect of Russia's nuclear command-and-control system malfunctioned, either

due to an internal bug or a malicious attack, then the USA stands to benefit if Russia's NC3 and NC4ISR systems are technically and procedurally up to the international standard of best cybersecurity practices. This is especially true in the face of heighted threats from a wider range of actors, assisted by the easy and rapid proliferation of weaponizable autonomy and machine learning techniques in the cyber domain.[32]

The shared interest in maintaining credible nuclear deterrent status therefore encourages norms to achieve best cybersecurity practices, which include avoiding integration of autonomy and machine learning in NC3 and NC4ISR systems and sharing provably secure digital platforms.

III. Conclusions

The introduction of autonomy and machine learning currently cannot be achieved without introducing new vulnerabilities that undermine the always/never alert status and the credibility of nuclear deterrence. Therefore, their integration into NC3 and NC4ISR systems should be avoided, for example through strict guidelines embedded in procurement policies.

In addition to unilateral action that can be taken by nuclear-armed states to reduce vulnerabilities and prepare for attacks, a second method for not increasing existential risk already posed by intentional, inadvertent or accidental nuclear war is to develop and institutionalize international norms and coordination mechanisms. There are three domains of particular relevance: (*a*) establishing an international norm prohibiting targeting of NC4ISR systems and nuclear weapon personnel, (*b*) promoting a norm against the weaponization of autonomy and machine learning, especially in the domains of cyberattacks and influence campaigns, and (*c*) sharing cybersecurity best practices and cyber-defences among nuclear-armed states, including the best practice of not integrating autonomy and machine learning into NC4ISR systems.

[32] Brundage et al. (note 14).

Shahar Avin 21
Deleted: 349