REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
04-06-2019	MASTER'S THESIS	

4. TITLE AND SUBTITLE

Machine Learning Systems in Nuclear Command, Control, and Communications Architecture: Opportunities, Limitations, and Recommendations for Strategic Commanders

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Falcone, Johnathan D., LT, USN

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Civilian Institutions Office (Code 522)
Naval Postgraduate School
1 University Circle, Herrmann Hall Rm HE046
Monterey, CA 93943-5033

10. SPONSOR/MONITOR'S ACRONYM(S)

NPS CIVINS

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

AI systems today and in the near future offer the potential to ease the time, uncertainty, and information deluge burdens that characterize NC2. At the same time, this emerging enabling technology is both limited and carries risk in its current technological state. The consequences of misapplication or error could be existential. As commanders look for ways to integrate these systems into their teams, they must do so deliberately and purposefully. This paper will assess the opportunities and challenges to apply ML techniques in NC2 systems. The results of this analysis show that although adoption challenges may exist, the potential to reduce errors and increase decision-making time are significant. As such, the paper will outline how the technology can be best adopted. It will offer recommendations that spans from development to employment in an effort to manage the risk associated with integrating the new technology. Ultimately, this paper does not aim to deter leaders, but instead highlight possible integration challenges so that this incredibly capable technology does not become stigmatized due to misapplications.

15. SUBJECT TERMS

Artificial Intelligence, Management, Emerging Technology, Nuclear Command and Control, Nuclear Decision-Making, Weapons of Mass Destruction

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPONE NUMBER (*Include area code*)
U	U	U	UU	26	

THIS PAGE INTENTIONALLY LEFT BLANK

Johnathan Falcone

Machine Learning Systems in Nuclear Command, Control, and Communications Architecture:

Opportunities, Limitations, and Recommendations for Strategic Commanders

"With the increase in the complexity of warfare, the science of war is increasingly dependent upon human guidance. No matter how complicated it may become, war is always waged by men."

— *FM22-10, Department of the Army Field Manual,* Leadership (1951)

Introduction

Time-restricted, uncertain, and data-dependent. These characteristics of nuclear command and control underscore the challenges strategists and nuclear force commanders face. As a result, nuclear powers – including the United States, China, and Russia – are beginning to discuss opportunities to apply artificial intelligence (AI)[1] to nuclear command and control (NC2) operations.[2] The start of the dialogue reflects recent advances in neural networks, algorithmic design, and computing power that have enabled high-reliability AI systems to be developed and employed in national security settings as sensitive as nuclear control.

Although these systems offer opportunities to increase the speed of analysis and improve accuracy, the state of technology today is limited, and end-users are still grappling with its application. As commanders and decision-makers begin to integrate AI-driven

[1] In the Introduction the term "artificial intelligence" is used in its colloquial sense. In the technical section, this paper will differentiate between "true" artificial intelligence and machine learning systems – the latter which forms the majority of the agents described and referenced in this paper.

[2] This paper will focus on AI's nuclear potential in the United States, China, and Russia. For U.S.-based research: Geist, Edward, and Andrew J. Lohn. "How Might Artificial Intelligence Affect the Risk of Nuclear War?" *RAND Corporation*, 24 Apr. 2018, rand.org/pubs/perspectives/PE296.html.; For a Russian perspective: Stefanovich, Dmitry. "Artificial Intelligence and Nuclear Weapons." *RIAC*, 6 May 2019, russiancouncil.ru/en/analytics-and-comments/analytics/artificial-intelligence-and-nuclear-weapons/.; A U.S. analyst's perspective on Chinese nuclear posture: Saalman, Lora. "Fear of False Negatives: AI and China's Nuclear Posture." *Bulletin of the Atomic Scientists*, 5 Dec. 2018, thebulletin.org/2018/04/fear-of-false-negatives-ai-and-chinas-nuclear-posture/.

systems into their watch floors and command centers, they must be aware of how these systems operate as part of their overall team. The promise of these systems and today's geo-political environment suggests that states and militaries will move quickly to implement these technologies. This implementation needs to be done intelligently for not only our security, but the safety of the world.

AI systems today and in the near future offer the potential to ease the time, uncertainty, and information deluge burdens that characterize NC2. At the same time, this emerging enabling technology is both limited and carries risk in its current technological state. The consequences of misapplication or error could be existential. As commanders look for ways to integrate these systems into their teams, they must do so deliberately and purposefully. This paper will assess the opportunities and challenges to apply ML techniques in NC2 systems. The results of this analysis show that although adoption challenges may exist, the potential to reduce errors and increase decision-making time are significant. As such, the paper will outline how the technology can be best adopted. It will offer recommendations that spans from development to employment in an effort to manage the risk associated with integrating the new technology. Ultimately, this paper does not aim to deter leaders, but instead highlight possible integration challenges so that this incredibly capable technology does not become stigmatized due to misapplications.

This paper will focus on AI and machine learning (ML) systems based on the current state of technology. Bounding the discussion to the technology available today and in the near-term, helps provide proper grounding. Otherwise, there is risk that the conversation can drift and lose relevance to the present-day challenges.[3] This paper first focuses on the

[3] This paper will use the following words or phrases rather interchangeably: "ML system", "learning system", "computer learning agent", and "agent". When projecting future application of AI and ML based on today's technology, this paper assumes the absence of a major breakthrough in the research. An example of a "major breakthrough" would be a development on par with the discovery of neural nets and the onset of deep learning systems.

subject's technical dimension. It will describe artificial intelligence broadly, establish a working definition for our purposes, and then focus on an agent's learning element and its capacity to "evolve" from performance-based learning. The second section outlines the United States' nuclear command, control, and communication (NC3) architecture from open-source literature. The next section will analyze the potential vulnerabilities to these computer agents and learning systems. In this section, I will highlight three primary sources of vulnerability: data, algorithms, and the individual user. The final section offers recommendations to mitigate the potential effects of the aforementioned vulnerabilities, emphasizing efforts to bridge the Silicon Valley-Pentagon divide, an agile development process, and personnel diversity.

How Do Computer Agents Learn?

According to Stuart Russel and Peter Norvig, an agent – whether human, a rules-based system, or an AI system – is anything that takes in information from its environment through sensors and then acts upon that environment through actuators.[4] A computer agent – to include AI – consists of an agent program and architecture.[5] The architecture functions to push the perceived environmental inputs to the program, runs the program, and provides the chosen actions to the actuators. For example, a "simple reflex" agent only utilizes the current environmental input, makes a decision based on a "condition-action rule", and acts back upon its environment through its actuators (Figure 1).

[4] Russell, Stuart J., and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd ed., Prentice-Hall, 2009., pg. 34
[5] An agent program is designed to complete the agent's function, which runs from initial environmental percepts to actions. Percepts refer to the agent's perceptual inputs at any given instant. Russel and Norvig (2009), 34.

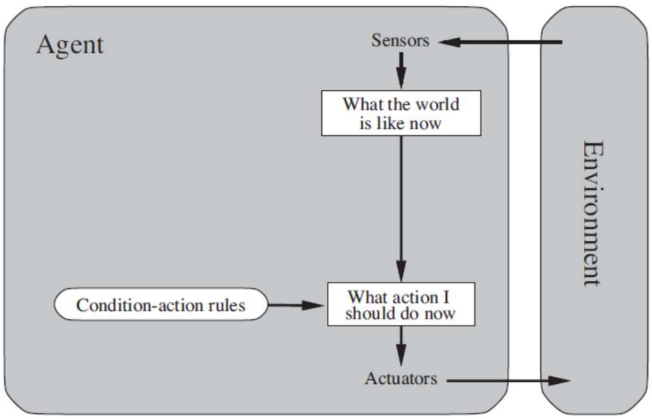

Figure 1, "Simple Reflex Agent"[6]

This agent type is inherently limited to its designers' imaginations and would be unable to operate in conditions not already anticipated. Even if all conditions could be anticipated, it would be burdensome to program all of these scenarios. Operations in complex environments such as warfare would be nearly impossible. As early as 1950, Alan Turing suggested that machines should be built to learn and then trained. Such a "learning" agent would offer significantly more potential. Today, these "learning" agents have come to supplant simple reflex – or rules-based – agents as the dominant agent type observed today.

Learning Agents

A learning agent is capable of learning from its own experiences. These experiences can be through interaction with its environment or training induced. This agent type is comprised of four general components: a learning element, performance element, critic, and problem generator (Figure 2). The first component, the learning element, is tasked with making improvements to the agent. Second, the performance element selects the external action. It is analogous to the entire simple-reflex agent in that this element takes inputs

[6] ibid, 49.

and delivers outputs through actuators. These first two elements are inextricably linked. The learning element can effect changes to the knowledge components after observing pairs of successive environmental states caused by the performance element's action. To facilitate this, the critic evaluates how the agent is performing in relation to a defined performance standard. The last general element is the problem generator which suggests exploratory actions that may undermine short-term performance for long-run effectiveness.[7]

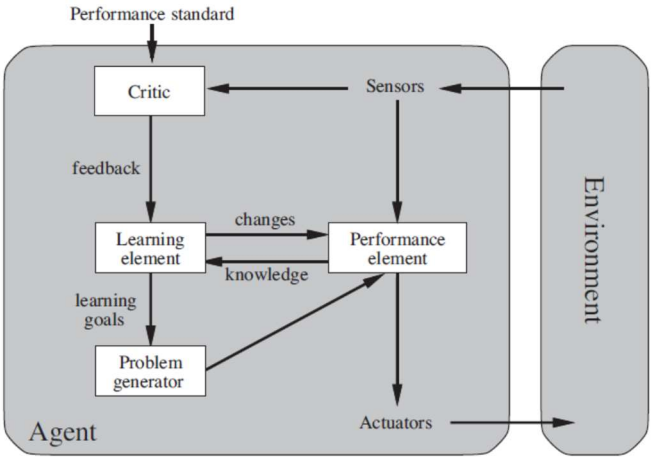

Figure 2: "Learning Agent"[8]

For the purposes of identifying these system's limitations and integration challenges in the NC3 architecture, it will be helpful to walk through these elements using a simple example. Take a learning agent developed for use in radiology. Its objective is to identify cancer images and produce a quantitative assessment from imaging data. This agent's task is to improve its ability to correctly diagnose patients. Its performance is to be measured with respect to the percentage of patients properly diagnosed based on its training with pre-diagnosed medical histories of patients.[9] Mechanically, sensors would observe images with

[7] ibid, 55-56.
[8] Russel and Norvig (2009), 55.
[9] Specifically, this example is illustrating a type of supervised learning process. Learning agents can also learn through unsupervised and reinforcement learning techniques.

pre-labeled features that were extracted and identified by human experts. The learning element will quantify these features based on the critic component's feedback, which is determined by the expert's performance standard. The problem generator will explore other characterizations and feed this information to the performance element. In the end, the performance element will classify the image quantifiably and provide this output to support human analysis in detection.[10] Each successive interaction with its environment – a new image – will be quantified into an attribute set and evaluated. The more interaction that an agent has with its environment, the more reliable it will become. In radiology, this type of learning helps decrease the false negative rate – the rate at which cancer goes unidentified from an image.

Machine Learning Today

Advances in the field have given birth to more efficient and effective learning agents. As opposed to "condition-action" rules, agents "learn" from algorithms and datasets. This approach is called machine learning and has been successful in enabling computer agents to complete specific tasks. Operationally, human-designed algorithms map features observed from the machine learning agent's environment and choose an output. Although commonly used interchangeably, machine learning and deep learning refer to two different categories within the AI field. In short, deep learning is a training technique within the machine learning subfield of artificial intelligence.

Deep learning is a specific machine learning technique that relies on artificial neural networks. Neural networks consist of many layers of artificial neurons which loosely mimic biological neurons. The network starts off as a blank slate, and algorithms "learn directly

[10] Hosny, Ahmed, et al. "Artificial Intelligence in Radiology." *U.S. National Library of Medicine*, Aug. 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6268174/#BX1.

by navigating the data space, giving them superior problem-solving capabilities."[11] Rather than identifying features utilizing a human-designed algorithm, deep learning systems extract features from raw data. Successive neural layers enable observations from an input image to be filtered increasingly more specifically. As data is extracted and selected through "hidden layers", pooling layers will aggregate the features that have been observed, and fully connected layers will enable classification. [12] Training a high-reliability system requires an appropriate number of hidden layers and large sets of data.[13]

These distinctions – between artificial intelligence, machine learning, and deep learning – are helpful when having discussions around these emerging technologies. Specifically, when we consider the limitations and challenges of adopting these systems into our operations, precision of language helps illuminate the actual issues at-hand.

An Enabling Technology

The last technical point for consideration before a broader discussion is to consider how AI and ML[14] will be used in NC3. AI and ML, as technologies, are not tools in and of themselves. Rather, they are merely an enabling technology similar to the combustion engine. The impact AI and ML will have on crisis management operations will be determined by their use within systems rather than their existence. Because their utility will be derived from their application, successful implementation will be the responsibility of all team members and require a new form of "adoption management."

[11] Hosny (2018).
[12] ibid
[13] The amount of data that is required for a machine learning algorithm to learn a target function is referred to as "sample complexity" in AI research.
[14] In the previous section I have defined that the state of today's technology is more accurately described by the term "machine learning". From this point forward, I will refer to systems in our current day as "ML", "learning systems", and "learning agents", which will be used to encompass both ML and deep learning systems. References to AI will be to the long-term actualization of autonomous, adaptable systems.

NC3 Architecture and Opportunities to Leverage Machine Learning Systems

The Nuclear Matters Handbook, published by the Office of the Deputy Assistant Secretary of Defense for Nuclear Matters, describes the U.S. Nuclear Command and Control System (NCCS)[15] as "an essential element to ensure crisis stability, deter attack against the United States and its allies, and maintain the safety, security, and effectiveness of the U.S. nuclear deterrent."[16] The system is expansive and includes land-, air-, sea- and space-based components.[17] These components support the President's ability "to authorize the use of nuclear weapons in a crisis and to prevent unauthorized or accidental use." [18] Furthermore, NCCS "relies on a collection of activities, processes, and procedures performed by appropriate military commanders and support personnel that…allow for senior-level decisions on nuclear weapons employment."[19] Interagency personnel from across the Executive Branch – to include Department of Defense, Department of State, Department of Justice, among others – are stakeholders in the NCCS.

The NC3 architecture is a complex system but can be generally categorized into four components. These four components highlight the interconnected nature, scale of available data, and consequences of action emanating from the system. The White House, Pentagon, and U.S. Strategic Command (USSTRATCOM) sit on top of the structure and the intersection of sensors, networks, and kinetic platforms. The primary fixed, integrative NC2 facility is the National Military Command Center (NMCC) inside the Pentagon.

[15] NCCS refers to the overall system that provides the President the means to authorize the use of nuclear weapons. NC3 refers to the architecture that includes command, control, and communications. NC2 refers strictly to the nuclear command and control.
[16] *2016 Nuclear Matters Handbook: Authoritative Guide to American Atomic Weapons, History, Testing, Safety, Security, Delivery Systems, Physics and Bomb Designs, Terror Threats*, 2017, https://www.acq.osd.mil/ncbdp/nm/NMHB/chapters/chapter_6.htm, 73.
[17] "Nuclear Command, Control, and Communications: Update on Air Force Oversight Effort and Selected Acquisition Programs" *U.S. Government Accountability Office*, 15 Aug. 2017.
[18] *2016 Nuclear Matters Handbook (*2017), 73.
[19] ibid, 73.

USSTRATCOM's Global Operation Center (GOC)[20] serves as an additional NC2 fixed-site facility. If the two fixed sites were unable to function, there are two airborne, survivable alternatives.

The Integrated Tactical Warning/Attack Assessment System (ITW/AA) are sensors that collect information on behalf of NCCS operations. These sensors include radars, infrared satellites, fixed and mobile sensors, and processing systems. The ITW/AA provides "unambiguous, reliable, accurate, timely, survivable, and enduring warning information of ballistic missile, space, and air attacks on North America."[21] The sensor network not only provides continuous surveillance, but also enables correlation and independent verification of potential threats. ITW/AA also consists of capabilities to warn the President, Secretary of Defense, and Chairman of the Joint Chiefs of Staff, and the assessment function reports whether a threat holds North America or U.S. assets at-risk.[22]

The transport component ensures both data and decisions can move reliably from sensors to command centers to delivery platforms. This necessitates robust, reliable, and survivable – including during a nuclear environment – communication methods. Additionally, NC3 relies on airborne relay and satellites to transmit and receive critical information. It also maintains land-based phone lines and undersea cables. If crisis or the need arises, NCCS must also be capable of transmitting commands and directives from decision-makers to all nuclear-capable platforms.

The fourth NC3 component is the nuclear weapon platforms and delivery systems capable of kinetic action. Air-launched platforms include heavy bombers[23] and dual-

[20] Located at Offutt Air Force Base near Omaha, Nebraska.
[21] ibid, 76.
[22] ibid, 73-76.
[23] Heavy bombers include the B-52H and B-2 which would deploy the W80-1 Air-Launched Cruise Missile (ALCM) and B61-7/11, B83-1 respectively. ibid, 25.

capable[24] aircraft. The Minuteman III system is the U.S. ground-launched delivery system and the Ohio-Class Ballistic Missile Submarines are America's sea-launched delivery systems.

Each of these platforms are capable of receiving, analyzing, and communicating data quickly. Threats to any of these systems – from communication lines to delivery platforms – that would reduce the system's ability to process data and disseminate information could be interpreted as a threat to the nuclear stability. The loss of data sources or communication channels would impose a limitation on the President's ability to make an informed decision and the system to execute that decision. As such, from the military's perspective, quickly identifying, managing, and holding these threats at-risk is critical. Systems that employ machine learning – and eventually true artificial intelligence – can aid in this task and contribute generally to NCCS objectives.

Reduce Errors, Maximize Capabilities, and Minimize Losses

The uses for ML systems in the NC3 context are vast. To limit the scope of this discussion, this paper will make continued use of NCCS objectives as defined in the 2016 Nuclear Matters Handbook. Using these objectives as a guidon, below are a few identified opportunities for ML – and eventually true AI – systems.

In the short term, one of the more natural uses of a deep learning system is with respect to automatic target recognition (ATR) to identify false positive sensor data to mitigate the chance of a tech-induced crisis. As previously discussed in the radiology example, neural network developments have contributed to great improvements in image recognition to reduce false negative diagnoses. Related to ATR, a false negative result

[24] Dual-capable aircraft can fulfill conventional or nuclear theatre missions. These platforms include F-15, F-16, some NATO aircraft, and the F-35. ibid, 25.

translates into an inbound missile that was either undetected, incorrectly classified as noise, or marked invalid. A false *positive* result, on the other hand, would be the appearance of an inbound track that did not exist in reality or was incorrectly classified as a missile. The objective of this type of learning agent would be to reduce the number of false positives – of which there are infamous cases[25] – and increase confidence with regards to negative outputs. Credibly reducing the false positive tracks will encourage more effective interrogation of the tracks that do appear. Anecdotally, when the system continually reports false positives the watch standers analyze those tracks less intensely. In turn, this increases the likelihood of false *negative* reporting. ML systems that reduce false positives will also likely increase watch stander vigilance, providing a substantive improvement to NC3 operations.

Computer agents can also be employed to rapidly synthesize intelligence, surveillance, and reconnaissance data from sensors to deter an attack against the United States and its allies. Today mobile launchers are difficult to track and target. Adversarial nations see advantages in road or mobile launchers because they can be moved regularly and striking them would require multiple missiles.[26] America's deterrent credibility can be strengthened by an agent with the ability to synthesize data related to an adversary's mobile launchers – including location and capability – and helps hold them at-risk. The threat posed by foreign mobile launchers would be reduced if a system had the ability to collate continuously updated sensor data with regards to the disposition of mobile launchers.

[25] In September 1983, Soviet Lt. Col. Stanislav Petrov chose to reject a "launch" command from a Soviet early warning system when it was reported with "high reliability" that an American ICBM was inbound. Matthews, Dylan. "35 Years Ago Today, One Man Saved Us from World-Ending Nuclear War." Vox, Vox, 26 Sept. 2018, www.vox.com/2018/9/26/17905796/nuclear-war-1983-stanislav-petrov-soviet-union.

[26] Beckhusen, Robert. "American Mobile Nuclear Missile Launchers Is a Really Bad Idea." The National Interest, 3 Oct. 2017, nationalinterest.org/blog/the-buzz/american-mobile-nuclear-missile-launchers-really-bad-idea-22579.

The security and effectiveness of the U.S. nuclear deterrent could also be aided by a force management / decision support system that maximizes second-strike capability. Deterrence theory suggests that for a deterrent to be effective, it must credibly dissuade adversaries from action through the certainty of retaliation. In this vein, the Nuclear Matters Handbook states that a second-strike capability is critical to achieving deterrence against an adversary.[27] The decision support system could draw from the inventory of available sensors to analyze an expansive variable list – including adversary force disposition, current U.S. force laydown, nuclear and conventional capabilities of both states, damage prediction, etc. – to produce an output that would aid the President and senior advisors in decision making.

It is also important to acknowledge that each of these systems have the potential to upset strategic stability and could exacerbate the security dilemma. It would be difficult for adversaries to tell the difference between capabilities that are clarifying unknowns versus those systems contributing to the capability for preemptive, first-use strike.

Vulnerabilities, Limitations, and Misuse

As described above, ML has vast potential to improve decision-making and minimize errors throughout the NC3 architecture. In addition to recognizing these opportunities, commanders and mid-grade leadership must also be aware of the technical and psychological limitations of adopting ML systems and integrating them into their teams. In an earlier discussion, this paper identified that learning systems require effective algorithms and training data. The suggestion that these technical inputs must be of a certain standard also suggests that these are system vulnerabilities. Additionally, human

[27] *2016 Nuclear Matters Handbook* (2017), 15.

users bring their own psychological limitations to the tech-adoption process. As an enabling technology and decision-making aid, the ultimate purpose of integrating these systems is to assist people. Looking to research in the psychology field, leadership should be aware that users are liable to both over rely and underutilize these applications because of human biases and heuristics.

Sources of Error: Data

Stated generally, learning systems are only as good as the data that trains them. A McKinsey Global Institute report describes data as "fuel" and states that "AI[28] cannot run without a steady diet of data." Machine learning systems require access to data sets to discover patterns, make associations, and develop insights.[29] A natural deduction is that learning systems are vulnerable to "bad" data. In practice, subversive data can be from intentional or unintentional sources.

During the development of any warfare technology it is reasonable to assume that an adversary is going to seek out countertactics or counterstrategy. In the case of data-trained learning systems, adversaries may look to deploy subversive data techniques. Two types of subversive techniques are "data poisoning"[30] and "adversarial examples"[31]. In both cases, these tactics target machine learning system's susceptibility to learn incorrectly or misidentify information. A "data poisoning" tactic would "inject false training data with the

[28] This quotation uses AI as a generic term.
[29] Barton, Dominic, et al. "Artificial Intelligence: Implications for China." *McKinsey Global Institute*, 2018.
[30] Steinhardt, Jacob, et al. "Certified Defenses for Data Poisoning Attacks." *31st Conference on Neural Information Processing Systems*, 2017.
[31] Goodfellow, Ian. "Deep Learning Adversarial Examples – Clarifying Misconceptions." *KDnuggets Analytics*, July 2015, www.kdnuggets.com/2015/07/deep-learning-adversarial-examples-misconceptions.html.

aim of corrupting the learned model."[32] This may be achieved through data security breaches or statistical manipulation techniques.[33]

"Adversarial examples" seek to fool learning agents. In a relatively mundane application, Cornell researchers demonstrated that "state-of-the-art" deep neural network image recognition systems could be tricked with "white noise". In their study, a system that was trained to label images was presented a "static" image and the system misidentified it as a lion with 99.99% confidence.[34] In the NC3 context, this susceptibility to "fooling" – sometimes referred to as spoofing – could result in false positive tracks. This could trigger a system alert that reports an attack is imminent. In reality it could be the use of an adversarial image that is imperceptible to human senses. A false positive track from a ML system is particularly concerning because human users may be biased to believe the system's output. This phenomenon is called "automation bias" and will be discussed in a future section. From a technical perspective, "there is no known way of fully inoculating algorithms against these attacks."[35] Adversaries do not need access to the underlying data or algorithm to take advantage of this vulnerability, which makes this data tactic more challenging to defend against. Deep learning systems are also susceptible to this vulnerability.[36]

The threats to data are not just limited to an adversary's attempts to hack machine learning systems. Unintentional data subversions could also be detrimental. This might be particularly true when introducing these systems into NC3. Given the lack of real-world

[32] Steinhardt (2017).
[33] On a spectrum of complexity, data security breaches would be at the low-end of data poisoning techniques while statistical manipulation is still in theory only. Security breaches are discussed in the Steinhardt papers. The statistical manipulation techniques are based on conversations the author has had with DoD AI-Consultants.
[34] In another example, an imperceptible change of the image tricked the system to relabel a lion's image to "library". Nguyen, Anh, et al. "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images." *Cornell University*, 2 Apr. 2015.
[35] Scharre, Paul. "Killer Apps." *Foreign Affairs*, 1 May 2019, www.foreignaffairs.com/articles/2019-04-16/killer-apps.
[36] ibid

scenarios involving nuclear crisis operations, data is extremely limited.[37] Therefore, learning systems that are employed in the NC3 field will likely be trained utilizing wargame data. Wargame data may contain three potential sources of data bias. The first is the assumption that players will make the same decisions in real-life as they did in a simulation. Although simulations are generally effective at reflecting real-world conditions, it may not be possible to truly simulate the moral weight that surrounds deciding to use nuclear weapons.

The second, that strategic decision making in wargames will naturally be based on the responses of a "red team" within the game. Can we ever be confident that the "red team" is accurately reflecting the most likely strategic decisions of an adversary? To say yes, makes two assumptions – that red team participants can accurately reflect adversarial strategy and tactics, and that they are allowed to do so. First, it would be presumptuous for any "red team" to believe they could effectively deploy their knowledge of Chinese or Russian strategic thinking as it relates to nuclear weapons in a wargame. There is minimal precedent and understanding to make those assertions. Second, provided new evidence, would participants be empowered to challenge the status quo perception of adversary decision-making? For example, Millennium Challenge 2002 was a U.S.-led wargame that simulated a war in the Middle East. In the games initial trial, Lt. Gen Paul Van Riper had decimated the American forces using "unconventional" tactics.[38] Lt. Gen. Van Riper was chastised for performing irrationally and when the wargame was reset for a second round,

[37] I do not believe I am in a position to claim that this data does not exist. There are most likely case studies – from the Cold War era – where tactical or strategic nuclear options were discussed. Additionally, the Trump Administration's Nuclear Posture Review stated its intention to expand the U.S. nuclear arsenal to include low-yield tactical nuclear weapons. Wargames or actual events – data opportunities – likely suggested that there would be benefit to owning these weapons.
[38] Galloway, Joe. "Rumsfeld's War Games." Military.com, 26 Apr. 2006, web.archive.org/web/20060504005348/http://www.military.com/opinion/0%2C15202%2C95496%2C00.html.

the exercise was "almost entirely scripted to ensure a [US] win."[39] This story helps illustrate the difficulties of challenging perceptions of senior officers and officials within the military.

Finally, based on information available, the President has not participated in past strategic wargames.[40] Instead of his direct participation, someone has simulated the President's decision-making authority. In a recent analysis of Cold War wargames, it was found that wargamers used nuclear weapons more frequently in the game than the President has done in reality.[41] As a result, wargame data may inherently contain an "upwardly" biased dataset.[42] The biases that may be introduced from wargame data suggests that prior to the employment of any learning system trained using data, a discussion related to the possible direction of bias and a plan to mitigate these impacts should be thoroughly had.

Sources of Error: Defining and Controlling Learning

At their core, machine learning algorithms are the development and employment of mathematical and statistical models.[43] Biases in algorithmic design can stem from misarticulated problem framing. "Second-strike capability", for example, is a vague objective. To concretize this aim, the system may be designed to maximize second-strike capability vis-à-vis military infrastructure or against population centers. These

[39] Borger, Julian. "War Game Was Fixed to Ensure American Victory, Claims General." *The Guardian*, 21 Aug. 2002, www.theguardian.com/world/2002/aug/21/usa.julianborger.
[40] Thomas Schelling advised against presidential participation, "…I do not think he should ever be put in the position where people watch him and what he would do in a crisis." Pauly, Reid B.C. "Would U.S. Leaders Push the Button?" *International Security*, 43:2 Fall 2018, pp. 151-192.
[41] In Pauly's analysis of historical wargame data, nuclear weapons were used twice in twenty-six sample strategic wargames. Continued iterations of wargames would likely result in nuclear use. Even if it is incredibly rare, any result greater than 0 – the actual number – would be reflected by an "upward" bias in the system. Pauly (2018).
[42] "Upward" bias translates to "a greater likelihood" of nuclear use.
[43] Hao, Karen. "What Is Machine Learning? We Drew You Another Flowchart." MIT Technology Review, MIT Technology Review, 17 Apr. 2019, www.technologyreview.com/s/612437/what-is-machine-learning-we-drew-you-another-flowchart/.

counterforce versus countervalue strategies may both achieve the second-strike objective but may not meet the intentions of different strategic planners and administrations.

Additionally, choices made for what data the algorithm should consider during the data preparation can also alter the model's accuracy.[44] For example, currently China's People's Liberation Army Navy is growing their capacity for a more dispersed nuclear force.[45] If the aforementioned machine learning system to maximize second-strike capability were being designed today, should it account for the number of nuclear-ballistic missile capable submarines not imaged in-port? If these unlocated submarines do not have functional nuclear strike capability the model might upwardly bias its perception of China's threat. On the other hand, if these are not included, but development of this capability occurs sooner than our intelligence estimates predict, then the model might be downward biased.

The current state of the technology also results in algorithms learning in unexpected or uncontrolled ways. These uncontrolled evolutions occur when "we specify the wrong objective function, are not careful about the learning process, or commit other machine learning-related implementation."[46] These "unexpected results often result from evolution thwarting a researcher's intentions: by exploiting a bug in the code, by optimizing an uninteresting feature, or by failing to answer the intended research question."[47] Control issues highlight the parts of "alchemy" that currently characterize ML. That is to say, machines are learning, but the developer struggles to explain exactly how an algorithm

[44] Hao, Karen. "This Is How AI Bias Really Happens-and Why It's so Hard to Fix." MIT Technology Review, 4 Feb. 2019, www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.
[45] Yeo, Mike. "China Completing More Ballistic Missile Subs, with Plans for a New Version." *Defense News*, 6 May 2019, www.defensenews.com/global/asia-pacific/2019/05/06/china-completing-more-ballistic-missile-subs-with-plans-for-a-new-version/.
[46] Amodei, Dario, et al. *Concrete Problems in AI Safety*, 25 Jul. 2016, https://arxiv.org/pdf/1606.06565.pdf, 1-2.
[47] Lehman, Joel, et al. *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities*, 14 Aug. 2018, https://arxiv.org/pdf/1803.03453.pdf, 3.

evolves. Comical examples of this include a Tetris-playing robot that learned to pause the game rather than lose.[48] Or a robot that learned to walk on its elbows when programmers created an environment where all six of its legs were not functional.[49] These evolutions would be much less welcomed in the nuclear command and control structure, and right now developers are not able to track how these evolutions take place.

Sources of Error: Human Users

Researchers from the Wharton School at the University of Pennsylvania have coined the term "algorithm aversion". This refers to people's reluctance to trust the work of machines because they believe humans can perform better. This phenomenon was captured by researchers whose studies concluded "seeing algorithms err [made] people less confident in them and less likely to choose them over an inferior human forecaster."[50] According to the studies' results, "it seems that the errors that we tolerate in humans become less tolerable when machines make them."[51] This trust gap can result in human users being unwilling to use a system or to use it for its intended purposes.

The opposite of the trust gap is referred to as automation bias. According to Dr. Linda Skitka, a psychologist who studies the types of errors people are prone to make in automated decision-making environments, automation bias is the "use of automation as a heuristic replacement for vigilant information seeking and processing."[52] Rather than conducting their own investigation of available data, people are over-reliant on the solution presented to them by the computer. This bias can result in both errors of commission and

[48] Biggs, John. "Programmer Creates An AI To (Not Quite) Beat NES Games." *TechCrunch*, 14 Apr. 2013, techcrunch.com/2013/04/14/nes-robot/.
[49] Lehman (2018), pg. 14
[50] Dietvorst, Berkeley J., et al. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology*, vol. 144, no. 1, 2015, pp. 114–126., doi:10.1037/xge0000033.
[51] Dietvorst (2015), 115.
[52] Skitka, Linda J., et al. "Automation Bias and Errors: Are Crews Better Than Individuals?" The International Journal of Aviation Psychology, vol. 10, no. 1, 2000, pp. 85–97., doi:10.1207/s15327108ijap1001_5.

omission. Errors of commission are actions that the human actor takes that he or she should not, while errors of omission are those that were not taken that should have been. In Skitka's research, it appears that automation bias reinforced the latter of these error types. It is likely that this bias may be amplified in crises due to limited time and increased stress. To mitigate the impact of this bias, leadership must ensure that personnel are being trained to be critical consumers of data.

RAND conducted workshops to discuss the potential impact of advanced AI[53] on nuclear security in mid-2017. The human end-user came up in discussion related to whether humans would be comfortable or not allowing a computer to influence decision-making in a nuclear war scenario. Interestingly, although participants took positions on both sides of the debate, it appeared opinions were divided inter-generationally. Namely, younger participants felt more comfortable with the idea that AI-systems could influence decisions of this magnitude.

Recommendations to Integrate Machine Learning Systems in NC3

As an enabling technology, AI and ML systems are a revolutionary technology that will expand the possibilities frontier across the national security spectrum. Specific to the nuclear command and control field, it will help collate information and improve the rate of false negatives among the other opportunities previously discussed. All of these potentialities require that, despite limitations, strategic command leadership must adopt this technology. This adoption must be done carefully. If improperly applied, there is the risk that an entire generation of leaders will be reluctant to use the technology, fulfilling the algorithmic aversion bias. In an effort to bring these systems online in a productive

[53] For clarity, this RAND workshop was referencing true AI-capable systems. The projected timeline in the workshop was 2040.

manner, I offer three recommendations: One, bridge the civilian-military divide between Silicon Valley and the Pentagon through a military ambassadorial team stationed in Silicon Valley. Two, procure ML systems through an agile system procurement method. And three, ensure that the teams developing and utilizing these systems are diverse.

Build Civ-Mil Cooperation with Military Liaisons

One of the first challenges that will have to be overcome in order to procure ML agents for military use is to get over some of the reluctance that the most capable developers have demonstrated recently.[54] This rift has been attributed to moral objections from technologists and workers in the valley.[55] Others have argued that the divide is more practical and is driven by fiscal responsibility at a start-up.[56] Unlike other military technologies – such as missile development – large defense contractors do not have the expertise to develop systems at a level than can compete in the current global race.[57] This expertise and experience with ML – and eventually true AI systems –exists outside the traditional military-industrial complex. Thus, in an effort to build systems that are not burdened by some of the aforementioned limitations, it will be required for the two parties to talk to each other.

Past initiatives by the Pentagon to leverage the idea generation occurring in the start-up tech world include In-Q-Tel and Defense Innovation Unit. IQT is a not-for-profit strategic investor that funds startups who are creating innovative technologies for use in the national security space. DIU is a government entity that contracts with technology

[54] Tiku, Nitasha. "Why Tech Worker Dissent Is Going Viral." *Wired*, 29 June 2018, www.wired.com/story/why-tech-worker-dissent-is-going-viral/.
[55] O'Mara, Margaret. "Silicon Valley Can't Escape the Business of War." *The New York Times*, 26 Oct. 2018, www.nytimes.com/2018/10/26/opinion/amazon-bezos-pentagon-hq2.html.
[56] Olney, Rachel. "The Rift Between Silicon Valley and the Pentagon Is Economic, Not Moral." *War on the Rocks*, 6 Feb. 2019, warontherocks.com/2019/01/the-rift-between-silicon-valley-and-the-pentagon-is-economic-not-moral/.
[57] Fryer-Biggs, Zachary. "Inside the Pentagon's Plan to Win Over Silicon Valley." Wired, 21 Dec. 2018, www.wired.com/story/inside-the-pentagons-plan-to-win-over-silicon-valleys-ai-experts/.

companies to solve defense issues. Both of these entities positively contribute to innovation and procuring national security needs, but frankly, their scale is too small, and they do not address the interpersonal challenges between the Pentagon and Silicon Valley.

One means of closing this interpersonal gap would be through a military liaison unit posted in Silicon Valley. This recommendation is inspired by Denmark's Tech Ambassador who maintains a physical presence there. Out of recognition of "the key role that technology and digitalisation plays and will increasingly play in the future for individuals and societies alike," the Tech Ambassador contributes to a "stronger multi-stakeholder discussion on how we want these new technologies to shape our societies in the future."[58] Similarly, the military liaison unit would facilitate stakeholder discussions over key national security issues. This unit should be led by a 3-star General or Admiral reporting to CYBERCOM and the Deputy Under Secretary of Defense for Acquisition and Sustainment. After establishing the channel for communication between the Pentagon and Silicon Valley, the next step is to define how cooperation between the two parties should be conducted.

System Procurement Utilizing an Agile Process

In general, the two most fundamental forms of software development are traditional and agile approach. Traditional project management – also referred to as the waterfall approach – is a linear-phased development approach. According to the Project Management Institute, the waterfall methodology is characterized by a detailed, long-term project plan with a single timeline and a product that is delivered at the end of that timeline. The project management structure is more rigid than the agile method, and changes to the

[58] Office of Denmark's Tech Ambassador. "TechPlomacy." Office of Denmark's Tech Ambassador, 2019, techamb.um.dk/en/Techplomacy/.

deliverables are costly.[59] On the other hand, agile projects are based on shorter timelines and multiple delivery dates. Although the product timeline is less definitive and delivered at functional stages, changes to the product are frequent and expected.[60] Additionally, the end-user is involved throughout the agile process, helping guide each iteration, while the traditional method only involves the end-user at the beginning and end of the project.

The agile method is driven by twelve principles.[61] Of note, the agile process "welcome[s] changing requirements, even late in development. Agile processes harness change for the customer's competitive advantage."[62] This methodology encourages – and necessitates – stakeholder involvement and transparency. It also has shorter product delivery timelines, accommodates needs that are in flux, and focuses on the user.

For a ML end-product to be delivered on behalf of NC3 operations, the implementation needs to be seamless. Because of the potential for biases – from data and algorithms – it will be important for end-users, rather than developers, to stress test and run scenarios using the equipment to identify where these systems may go awry. Additionally, strategic developments and capacity – although not constantly changing – occur frequently enough that iterative products might require tweaks to underlying algorithms. For these reasons, an agile development process has greater potential to mitigate the potential vulnerabilities that have been explored in this paper.

Personnel Diversity

This paper has attempted to demonstrate that despite all of the advances in technology, fundamentally, human actions and decisions are behind building and using the

[59] Fair, Jason. Agile versus Waterfall: approach is right for my ERP project? Paper presented at PMI® Global Congress 2012—EMEA, Marsailles, France. Newtown Square, PA: Project Management Institute, 2012.
[60] ibid
[61] "Principles Behind the Agile Manifesto." *Principles behind the Agile Manifesto*, agilemanifesto.org/principles.html.
[62] ibid

machine learning systems that will be deployed in an NC3 capacity. Because the human element is still so fundamental, it is critical to bring in diverse worldviews to develop these technologies. Silicon valley – and the technology sector more broadly – is infamously plagued by a lack of diversity.[63] In warfare and the national security space, America's willingness to take advantage of diversity has made the nation more secure.[64] As the United States continues to recognize and lean forward into developing AI for the NC3 architecture, it is essential that it pushes tech to diversify in an effort to solve the complex challenges it will face in development.

It is important for personnel diversity to exist at all levels as machine learning systems are attempted to be brought online. From technical developers to command center watch standers to on-the-ground responders who will use and be impacted by the new applications, broad cognitive perspectives and backgrounds are essential. It is commonly understood that diversity encourages greater cooperation and better problem-solving ability in teams. This is likely even more important when developing emerging technology for crisis management purposes. First, in conjunction with a previous recommendation, the agile process requires close interaction between developers and end-users, necessitating cooperation. Second, mitigating the challenges of algorithmic bias will require multiple perspectives and intuitions about how to solve problems. Finally, because humans will be employing this technology, their individual psychology will impact how they rely and use the outputs provided. If all individuals approach the problem in a similar manner, the individual cognitive biases will be compounded. Given how a focus on ensuring diversity

[63] Carson, Erin. "When It Comes to Diversity, Tech's Idealism Keeps Falling Short." *CNET*, 8 Dec. 2018, www.cnet.com/news/when-it-comes-to-diversity-techs-idealism-keeps-falling-short/.
[64] Fanning, Eric. "America's Diversity Is Our Army's Strength." *Army.mil*, 30 Sept. 2016, www.army.mil/article/174964/americas_diversity_is_our_armys_strength.

could reduce data, algorithmic, and cognitive biases, it appears to be a critical factor in developing the most capable systems.

Conclusion

Given the current state of AI and ML technology, it is important to remember that instruments of war are inherently built by and deployed on behalf of people. This suggests that a learning system's inherent capabilities are robust but limited, employment may be well-intended but misinformed, and its impact may be tactically efficient but strategically destabilizing. As these technologies are adopted and integrated into the NC3 architecture, leadership must strike a balance between employing the technology without overreliance and maintaining tempered faith in its potential.

An outstanding item that requires additional research is the impact of introducing this technology to nuclear stability. Based on its current formulation, this nuclear balance can be potentially threatened or improved. Threats to this stability may result from a new imposition on the security dilemma. The development of machine learning technology has vastly improved machine vision and signal processing, which has improved the quality of percepts and programmatic fusion within the computer agent.[65] This technology may enable autonomous underwater vehicles that may eventually be capable – or give the perception of capability – of identifying and holding at risk another nation's "survivable forces", threatening their second strike capability and increasing the possibility that one might threaten a first strike. To mitigate this risk, it may be in U.S. interests to declare a "no first-use" policy.

[65] Geist and Lohn (2018), 10.

Optimistically, there is the potential for these systems to reduce tension and bring greater assurance to strategic stability. A 2018 RAND "Security 2040" report suggests three primary ways that this could occur. One, AI and ML systems, after a period of introduction where their reliability increases and limitations become better understood, would have capabilities that are less fallible and error-prone than human alternatives. Two, "launch under attack" postures may be reduced because AI and ML systems would contribute to higher-reliability early warning systems, intelligence gathering, and analysis resulting in greater first-strike stability. Finally, "radical transparency" among nuclear states regarding advanced learning systems limit the space for miscalculation. In order to achieve any of these optimistic outcomes, diplomatic negotiations and compromises will be required.

Whereas this paper aimed to highlight and make recommendations related to the current state of machine learning technology, much greater research is required to highlight the potential limitations of future AI-driven systems. The algorithm "control issue" seems of particular importance as new leaps and capabilities are developed within this space. Before advances in AI research allow these systems to transition from decision support aids to onboard expert systems, researchers must be able to explain algorithmic evolution. The vulnerability to "adversarial images" is also open-ended question. From the machine-user perspective, bridging the trust gap described by algorithmic aversion will be important to ensure that systems are being properly employed.

As technical advancements are made and state modernize their nuclear capability, it is inevitable that systems replicating human intelligence – ranging from automated to autonomous to artificial intelligence systems – will be employed in nuclear command and control. The overall objective of integrating an enabling technology like machine learning into the nuclear command and control sphere should be to help decision-makers have ample

time and information to make decisions of such complexity and consequence. As such, despite the rise of various emerging technologies, human beings continue to be the forces behind war. The limitations that humans have may be mitigated, but not eliminated, through the use of technology. If the prospect of nuclear war is on the horizon, this is not a responsibility to be delegated to a computer agent. These must remain human decisions where values, morality, and the character of a nation should carry more influence than an algorithm.