sipri

THE IMPACT OF ARTIFICIAL INTELLIGENCE ON STRATEGIC STABILITY AND NUCLEAR RISK

Volume I
Euro-Atlantic Perspectives

EDITED BY VINCENT BOULANIN

May 2019

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE

SIPRI is an independent international institute dedicated to research into conflict, armaments, arms control and disarmament. Established in 1966, SIPRI provides data, analysis and recommendations, based on open sources, to policymakers, researchers, media and the interested public.

The Governing Board is not responsible for the views expressed in the publications of the Institute.

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE
Signalistgatan 9
SE-169 72 Solna, Sweden
Telephone: + 46 8 655 9700
Email: sipri@sipri.org
Internet: www.sipri.org

The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk

Volume I
Euro-Atlantic Perspectives

EDITED BY VINCENT BOULANIN

May 2019

Contents

Preface

The post-cold war global strategic landscape is currently in an extended process of being redrawn as a result of a number of different trends. Most importantly, the underlying dynamics of world power are shifting with the economic, political and strategic rise of China, the reassertion under President Vladimir Putin of a great power role for Russia, and the disenchantment expressed by the current United States' administration with the international institutions and arrangements the USA had a big hand in creating. As a result, a binary Russian–US nuclear rivalry, legacy of the old Russian–US confrontation, is being gradually replaced by regional nuclear rivalries and strategic triangles. As the arms control framework that the Soviet Union and the USA created at the end of the cold war disintegrates, the commitment of the states with the largest nuclear arsenals to pursue stability through arms control and potentially disarmament is in doubt to an unprecedented degree. On top of this comes the impact of new technological developments on armament dynamics. The world is undergoing a 'fourth industrial' revolution, characterized by rapid and converging advances in multiple technologies including artificial intelligence (AI), robotics, quantum technology, nanotechnology, biotechnology and digital fabrication. How these technologies will be utilized remains a question that has not yet been fully answered. It is beyond dispute, however, that nuclear-armed states will seek to leverage these technologies for their national security.

The potential impact of these developments on strategic stability and nuclear risk has not yet been systematically documented and analyzed. AI is deemed by many to be potentially the most transformative technology of the fourth industrial revolution. The SIPRI project, 'Mapping the impact of machine learning and autonomy on strategic stability,' is a first attempt to present a nuanced analysis of what impact the exploitation of AI could have global strategic landscape and whether and how it might undermine international security. This edited volume is the first major publication of this two-year research project; it will be followed by two more. The authors of this first volume are experts from Europe, Russia and the USA; the two succeeding volumes will bring together contributions from South Asian and East Asian experts. The result will be a wide-ranging compilation of regional perspectives on the impact that recent advances in AI could have on nuclear weapons and doctrines, strategic stability and nuclear risk.

SIPRI commends this study to decision makers in the realms of arms control, defense and foreign affairs, to researchers and students in departments of Politics, International Relations and Computer Science as well as members of the general public who have a professional and personal interest in the subject.

Dan Smith
Director, SIPRI
Stockholm, May 2019

Acknowledgements

I would like to express my sincere gratitude to the Carnegie Corporation of New York for its generous financial support of the project. I am also indebted to all the experts who participated in the workshop that SIPRI organized on the topic on 22–23 May 2018 and who agreed to contribute to this volume. The essays that follow are, by and large, more developed versions of the presentations delivered at the workshop, taking into account the points made in subsequent discussions. The mix of perspectives achieved at this workshop is reflected in the different styles and substance of the chapters. The views expressed by the various authors are their own and should not be taken to reflect the views of either SIPRI or the Carnegie Corporation of New York.

I wish to thank SIPRI colleagues Lora Saalman, Su Fei, Tytti Erästö and Sibylle Bauer for their comprehensive and constructive feedback. Finally, I would like to acknowledge the invaluable editorial work of David Cruickshank and the SIPRI Editorial Department.

Vincent Boulanin

Abbreviations

A2/AD	Anti-access/area-denial
AAV	Autonomous aerial vehicle
AGI	Artificial general intelligence
AI	Artificial intelligence
ASV	Autonomous surface vehicle
ATR	Automatic target recognition
BMD	Ballistic missile defence
C3I	Command, control, communications and intelligence
C4ISR	Command, control, communications, computers, intelligence, surveillance and reconnaissance
CCW	(Convention on) Certain Conventional Weapons
DARPA	Defense Advanced Research Projects Agency (USA)
DLT	Distributed ledger technology
DOD	Department of Defense (USA)
GAN	Generative adversarial network
GAO	Government Accountability Office (USA)
ICBM	Intercontinental ballistic missile
INF Treaty	Treaty on the Elimination of Intermediate-Range and Shorter-Range Missiles
ISR	Intelligence, surveillance and reconnaissance
LAWS	Lethal autonomous weapon systems
NATO	North Atlantic Treaty Organization
NC3	Nuclear command, control and communications
NC4ISR	Nuclear command, control, communications, computers, intelligence, surveillance and reconnaissance
NORAD	North American Aerospace Defense Command
NPT	Non-Proliferation Treaty
SACCS	Strategic Automated Command and Control System
SAGE	Semi-Automatic ground environment
SAM	Surface-to-air missile
SEAD/DEAD	Suppression or destruction of enemy air defences
SSBN	Nuclear-powered ballistic missile submarine
TCB	Trusted computing base
UAV	Unmanned aerial vehicle
UCAV	Unmanned combat aerial vehicle
UUV	Unmanned underwater vehicle

Executive Summary

Artificial intelligence (AI) is undergoing a major renaissance. Since the beginning of the decade, a breakthrough in machine learning—an approach to AI engineering—has enabled the development of increasingly capable AI applications and autonomous systems. In the military realm, advances have created many expectations but also concerns, be it from a legal; ethical; operational or strategic standpoint. This edited volume focuses on the latter concern: the impact on AI on nuclear strategy. It is the first instalment in a trilogy that explores regional perspectives and trends related to the impact that recent advances in AI could have for nuclear weapons and doctrines; strategic stability and nuclear risk. It assembles the perspectives of 14 experts from the Euro-Atlantic community on why and how machine learning and autonomy might become the focus of an arms race among nuclear-armed states; and how the adoption of these technologies might impact their calculation of strategic stability and nuclear risk at the regional level and trans-regional level.

As far as the risk picture is concerned, contributors to this volume generally reach similar conclusions. They broadly agree that it is easy to misperceive the opportunities and challenges posed by AI in the military sphere. AI could enable major improvements in many areas of warfare, including the nuclear domain; however, foreseeable developments will be far more prosaic than the common representation of AI in popular culture. Super-intelligent AI systems that can learn and teach themselves to resist human control or Terminator-like autonomous systems are not the type of technology policymakers and the general public should be worried about. Rather, the main issue is that the military might underestimate or disregard the limitations of current AI technology. Machine learning powered AI applications and autonomous systems can achieve great things but remain brittle in their design. They may fail spectacularly when confronted with tasks or environments that differ slightly to those they were trained for. Their behaviour is also unpredictable as they use algorithms that are opaque. It is difficult for humans to explain how they work and whether they include bias that could lead to problematic—if not dangerous—behaviours. They could also be defeated by an intelligent adversary through a cyberattack or even a simple sensor spoofing trick. An immature adoption of the latest developments of AI in the context of nuclear weapons systems could have dramatic consequences.

Most contributors stressed in that regard, it would be prudent for states to devote time and resources to develop a clearer understanding of the limitations of AI and how they can be mitigated. Some authors fear, however, that the potential benefits of rapid military adoption of the advances of AI may prove irresistible to some nuclear-armed states, which would opt to lower their safety systems and reliability standards in order to maintain or develop their technological edge over their competitors.

The inherent nature of AI technology is, in that regard, a source of the problem: it is a software-based technology that makes a tangible evaluation of military

capabilities difficult. Nuclear-armed states could therefore easily misperceive their adversaries' capabilities and intentions. In the field of nuclear strategy and deterrence, the perception of an enemy's capability matters as much as its actual capability. A worrisome scenario would be a situation where a nuclear-armed state would trigger destabilizing measures (e.g. adopting new and untested technology or changing its nuclear doctrine) based only on the belief that its retaliatory capacity could be defeated by another state's AI capabilities.

Contributors all agree that an international discussion needs to be had on the opportunities and risks posed by the militaries' use of AI particularly in the nuclear capability-related context. Some contributors also stressed that the discussion needs to be inclusive: it may start with—but is not limited to—a conversation between like-minded countries. North Atlantic Treaty Organization (NATO) member states, Russia and other nuclear-armed states such as China and India should be engaging with each other on this issue. This engagement can be bilaterally and multilaterally through existing arms control and nuclear risk reduction discussion tracks. Civil society organisations, academia and industry should also be invited to play a greater role in these interstate discussions, they can help states to better understand the developmental trajectory of technology and the associated risks.

In terms of options for dealing with the risks, a number of contributors acknowledged that the types of risks posed by AI in the nuclear domain are not necessarily new. Recent advances of AI exacerbate old and well-known risks rather create new ones; which means that the solutions for dealing with them already exist. No first use policies, lowering the alert status of nuclear arsenals as well as traditional approaches to transparency and information sharing could help. This does not mean that states should shy away from exploring new policy options. These could include legally binding agreements on the need to maintain human control over nuclear launch decisions or policy binding confidence- and security-building measures and agreements such as to not to use AI to actively interfere with command-and-control structures.

Introduction

1. Introduction

VINCENT BOULANIN

Since the beginning of this decade the field of artificial intelligence (AI) has been undergoing a major renaissance—particularly due to a breakthrough in machine learning that has enabled the development of increasingly capable AI applications and autonomous systems. This renaissance has raised many hopes but also concerns. On the one hand, some AI experts have compared the transformative potential of AI to that of electricity: 'Just as everything became more useful when it was "electrified", everything will become more useful when it is "cognified"'.[1] In the military realm, this means that AI could make any type of military system—whether cyber, conventional or nuclear—smarter or more autonomous. On the other hand, AI systems have a number of limitations that make their potential use problematic from ethical, legal and security perspectives. If not properly programmed or used, AI systems could misinform human decisions and actions (e.g. reinforce existing human bias or create new ones). They could also fail in unpredictable ways or be particularly vulnerable to cyberattacks. In the military context, the potential consequences of these limitations could be dramatic.[2]

Expectations and concerns associated with the military use of AI are the focus of a growing literature.[3] Much of it is connected to the ongoing intergovernmental discussion on lethal autonomous weapon systems (LAWS) that is taking place in the framework of the 1980 Convention on Certain Conventional Weapons (CCW Convention).[4] As its name suggests, the CCW Convention focuses on conventional weapons, not issues related to nuclear weapons and strategic

[1] Ng, A. cited in Kostopoulos, L., 'AI, emerging tech and national defense @ SIPRI Stockholm Security Conference', Medium, 23 Sep. 2018.

[2] Boulanin, V., 'Mapping the debate on LAWS at the CCW: taking stock and moving forward', EU Non-proliferation Paper no. 49, EU Non-proliferation Consortium, Mar. 2016.

[3] E.g. Allen, G. and Chan, T., *Artificial Intelligence and National Security* (Belfer Center for Science and International Affairs: Cambridge, MA, July 2017); Cummings, M. L., *Artificial Intelligence and the Future of Warfare* (Chatham House: London, Jan. 2017); Sharikov, P., 'Artificial intelligence, cyberattack, and nuclear weapons—a dangerous combination', *Bulletin of the Atomic Scientists*, vol. 74, no. 6 (2018), pp. 368–73; Heinl, C. H., 'Artificial (intelligent) agents and active cyber defence: policy implications', eds P. Brangetto, M. Maybaum and J. Stinissen, *2014 6th International Conference on Cyber Conflict: Proceedings* (IEEE: Piscataway, NJ, 2014), pp. 53–66; Schuller, A. L., 'At the crossroads of control: the intersection of artificial intelligence in autonomous weapon systems with international humanitarian law', *Harvard National Security Journal*, vol. 8, no. 2 (May 2017), pp. 379–425; De Spiegeleire, S., Maas, M. and Sweijs, T., *Artificial Intelligence and the Future of Defense: Strategic Implications for Small- and Medium-sized Force Providers* (Hague Centre for Strategic Studies: The Hague, 2017); and Defense One, *AI, Autonomy and the Future Battlefield* (Defense One: Washington, DC, Feb. 2017).

[4] Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects (CCW Convention), opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983. See also Asaro, P., 'On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making', *International Review of the Red Cross*, vol. 94, no. 886 (summer 2012), pp. 687–709; Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017); Marchant, G. E. et al., 'International governance of autonomous military robots', *Columbia Science and Technology Law Review*, vol. 12 (2015), pp. 272–315; and Horowitz, M. C., 'The ethics and morality of robotic warfare: assessing the debate over autonomous weapons', *Dædalus*, vol. 145, no. 4 (fall 2016), pp. 1–16.

Box 1.1. Key definitions

Artificial intelligence

Artificial intelligence is a catch-all term that refers to a wide set of computational techniques that allow computers and robots to solve complex, seemingly abstract problems that had previously yielded only to human cognition.[a]

Nuclear weapon systems

Nuclear weapon systems should be understood in the broadest sense. They include not only the nuclear warheads and the delivery systems but also all nuclear force-related systems such as nuclear command and control, early-warning systems and intelligence, reconnaissance and surveillance systems. Relevant non-nuclear strategic weapons include long-range high-precision missiles, unmanned combat aerial vehicles (UCAVs) and ballistic missile defence systems.

Strategic stability

Strategic stability has many definitions. It is understood here as 'a state of affairs in which countries are confident that their adversaries would not be able to undermine their nuclear deterrent capability' using nuclear, conventional, cyber or other unconventional means.[b]

[a] See the detailed definition in chapter 2 in this volume.
[b] Podvig, P., 'The myth of strategic stability', *Bulletin of the Atomic Scientists*, 31 Oct. 2012.

stability. However, the transformative potential of AI is also relevant for nuclear weapons and nuclear doctrines.[5] AI could even be a driver of great entanglement between nuclear and conventional weapons.[6] The impact of AI in the field of nuclear weapons and doctrines therefore deserves greater scrutiny.[7]

To support a conversation on this topic, SIPRI organized a series of regional workshops: in Stockholm, Sweden, in May 2018, in Beijing, China, in September 2018 and in Colombo, Sri Lanka, in February 2019. The purpose of this workshop series was threefold.

1. The workshops were to raise awareness among both AI experts and those scholars and practitioners who work on nuclear weapon-related issues of the impact that the current AI renaissance could have on nuclear weapons and doctrines and on strategic stability more generally (see box 1.1).

2. The three workshops were to facilitate a global conversation on the topic by allowing experts from various parts of the world to interact and share their regional or national perspectives.

[5] On the equivalent effect on biological weapons see Brockmann, K., Bauer, S. and Boulanin, V., *Bio Plus X: Arms Control and the Convergence of Biology and Emerging Technologies* (SIPRI: Stockholm, Mar. 2019).

[6] 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; and Acton, J. M. (ed.), *Entanglement: Chinese and Russian Perspectives on Non-nuclear Weapons and Nuclear Risks* (Carnegie Endowment for International Peace: Washington, DC, 2017).

[7] Notable studies exploring this issue among the sparse examples include Altmann, J. and Sauer, F., 'Autonomous weapon systems and strategic stability', *Survival*, vol. 59, no. 5 (Nov. 2017), pp. 117–42; Payne, K., 'Artificial intelligence: a revolution in strategic affairs?', *Survival*, vol. 60, no. 5 (Oct.–Nov. 2018), pp. 7–32; Horowitz, M. C. et al., *Strategic Competition in an Era of Artificial Intelligence* (Center for New American Security: Washington, DC, July 2018); Horowitz, M. C., 'Artificial intelligence, international competition, and the balance of power', *Texas National Security Review*, vol. 1, no. 3 (May 2018), pp. 37–57; Geist, E. and Lohn, A. J., *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (Rand Corporation: Santa Monica, CA, 2018); and Lieber, K. A. and Press, D. G., 'The new era of counterforce: technological change and the future of nuclear deterrence', *International Security*, vol. 41, no. 4 (spring 2017), pp. 9–49.

3. The workshop participants were to identify the impact that the interconnection between AI and nuclear weapons could have on strategic relations in different regions of the globe. One of the hypotheses underlying the workshop design was that the impact would be different in each region: the Euro-Atlantic region (i.e. Europe and North America), East Asia and South Asia.

This collection of essays is based on the proceedings of the first regional workshop, held in Stockholm in May 2018. This workshop was primarily intended to collect views from experts with complementary backgrounds from the Euro-Atlantic community (specifically, France, Germany, Norway, Russia, Switzerland, Sweden, the United Kingdom and the United States), but it also involved experts from elsewhere (i.e. China, India, Israel and Pakistan). The workshop consisted of a series of panel discussions, which explored different aspects of the topic, and two break-out sessions, in which participants analysed in smaller groups the risks that military applications of AI could pose to strategic stability and how those risks could be mitigated.

Overview

The essays in this volume are grouped into three thematic parts, each with a short introduction by the editor.

Demystifying artificial intelligence and its military implications

Part I aims to provide the reader with a nuanced picture of the state of AI and how its recent advances could or could not be exploited by the military in the near future.

The present author starts (in chapter 2) with a basic introduction to AI: how it is commonly defined, how AI approaches have varied over time and what is at stake with the current AI renaissance. The essay explains that it is machine learning—a specific approach to AI engineering—that is primarily responsible for the recent leaps forward in the development of AI applications, notably autonomous systems. The importance that machine learning and autonomous systems currently have for the field of AI is the primary rationale for the focus of this volume. Rather than discussing AI generally, the contributors to this volume have been tasked to discuss specifically the potential and impact that advances in these specific technologies could have in the military sphere, and particularly in the realm of nuclear weapons and doctrines.

In the second essay (chapter 3), Dimitri Scheftelowitsch, a computer scientist at TU Dortmund University, focuses on an application area that can be deemed to be a key by-product of the current AI renaissance: autonomous systems. He explains, from the perspective of an engineer, how autonomous systems work, what types of task state-of-the-art autonomous systems can undertake, and what the technical and security challenges related to the design and use of these systems are. Scheftelowitsch argues that, while recent breakthroughs in AI have made possible the automation of several tasks previously considered as complex (e.g. dependable vehicle control or air traffic control), there remain technical limits

on what computers and robots can achieve autonomously. He explains that, for many tasks and operating environments, the design of autonomous systems that can be used in practice remains a considerable engineering challenge because the computer power requirements are too significant or because the decision process is too difficult to model. There are also important technical issues that need to be solved in order to ensure that these systems can perform in a reliable and safe fashion once deployed, which is of paramount importance when the systems have military uses.

In the final essay of part I (chapter 4), Martin Hagström of the Swedish Defence Research Agency echoes the views of Scheftelowitsch on the state of machine learning and autonomous systems. Focusing on the case of military systems, he points out that, while automation has been used for almost a century in military systems, including weapon systems, there are many remaining challenges for wide-ranging generic applications of autonomy in the military sphere. One of the most notable—from a technical and operational perspective—is the paradoxical requirement for predictability. The systems' behaviour has to be predictable to the military operators that use them, but not predictable enough to an enemy who could otherwise exploit this predictability to its advantage. For Hagström, the recent breakthrough in machine learning, which has contributed to solving numerous problems in several fields of AI, could improve the design of autonomous military systems and offer major qualitative improvements to a large variety of military applications, from cyber-defence systems to information-management systems for intelligence, reconnaissance and surveillance (ISR). He explains, however, that use of machine learning algorithms further complicates predictability for the military, given that a key characteristic of models created by machine learning is that they are not transparent and so cannot be certified for use using existing methods of testing and verification. Machine learning as an engineering approach has, moreover, its own technical limitations. Training machine learning systems requires large amounts of pre-recorded or computer-simulated data representing the application, but many key military problems inherently lack data (i.e. are data-thin) and may not be easily captured in a computer simulation.

Artificial intelligence and nuclear weapons and doctrines: past, present and future

Part II explores the specific connections between AI and nuclear weapons and doctrines.

In the opening essay of this part (chapter 5), John Borrie of the United Nations Institute for Disarmament Research (UNIDIR) recalls that the connection between AI and nuclear weapons and doctrines is not new: nuclear-armed states saw as early as the 1960s that the nascent field of AI could play a role in the development and maintenance of their retaliatory capability. The Soviet Union and the United States pursued the development of AI systems that would make their command-and-control processes more automated, giving policymakers more time to make critical decisions. At the same time, as Borrie explains, nuclear policymakers in the USA and the USSR also saw that AI technologies had real limits that required meaningful human control and supervision.

The key question then is: what might change with the advances in machine learning capabilities and autonomous systems? This is the issue that the present author addresses in the next essay (chapter 6). The conclusion is that, while it can be established that these technological advances could, theoretically, be exploited in all aspects of the nuclear deterrence architecture (from early warning and ISR, via command and control to nuclear weapon delivery), such applications might not necessarily be game-changing. They will bring qualitative improvements but do not resolve some of the fundamental questions of what constitutes an appropriate role for autonomous systems and what is an appropriate level of delegation of assessment and decision-making to machines.

The final two essays in part II assess the role that machine learning and autonomy seem to currently play in Russian and US nuclear force modernization plans. Notably, Page Stoutland of the Nuclear Threat Initiative (NTI) and Petr Topychkanov of SIPRI have chosen radically different focuses for their essays (chapters 7 and 8): while Stoutland's essay on the USA is all about machine learning, Topychkanov's on Russia primarily discusses the role of autonomy in nuclear weapon systems. This difference could be explained by the fact that the two countries seem to have different approaches to the role of AI in nuclear decision-making. According to Stoutland, in the USA it is commonly agreed that a human must make the decision to use a nuclear weapon, while machine learning and autonomous systems can only have a supporting role. He warns, however, that even when restricted to a supporting role, the use of machine learning could have important nuclear policy implications. In contrast, according to Topychkanov, Russia decided in 2011 to reactivate the fully automated nuclear command-and-control systems that the USSR developed during the cold war for nuclear retaliation. He also reports that Russia is currently exploring the possibility of developing autonomous offensive systems that could potentially be capable of delivering nuclear weapons.

Artificial intelligence, strategic stability and nuclear risk: Euro-Atlantic perspectives

Part III investigates the impact that the current or potential incorporation of AI into nuclear force systems could have on strategic stability and nuclear risk from the various perspectives of the Euro-Atlantic expert community. Six scholars from both sides of the Atlantic and two high-level UN practitioners try to address the question of how and how much the current status quo between nuclear powers could be undermined by the adoption of systems based on AI by nuclear-armed states, be it for conventional or nuclear weapons. These essays are meant to provide an impression of the collective Euro-Atlantic perspective on AI and nuclear weapons, rather than specific national or official opinions. It is notable in that regard that that the contributors, regardless of their country of origin, seem to reach the same broad conclusion.

For Michael Horowitz of the University of Pennsylvania (in chapter 9), using AI to automate parts of nuclear command and control and some aspects of nuclear weapon delivery could have both positive and negative consequences

for nuclear stability. In some cases, AI could improve safety and reliability in nuclear operations, notably by providing decision makers with better information and more time to make decisions. But AI systems are also brittle: they are likely to fail in situations that were not predicted in the design phase, which could potentially lead to accidental or inadvertent nuclear escalation. However, the greatest risk, according to Horowitz, may perhaps come from the way in which the military could use AI-based applications and autonomous weapon systems to fight conventional wars at greater speed. Horowitz argues that the fear of being outpaced in the conventional realm could create incentives for nuclear-armed states that are not confident in their second-strike capability to adopt unstable nuclear postures such as launch on warning or even to strike first in a crisis.

Frank Sauer of Bunderswehr University Munich arrives at a similar conclusion (in chapter 10). He shares Horowitz's view that it is the conventional applications of AI that cause worry. He explains that military applications of AI and machine learning in conventional warfare generate a greater entanglement between the conventional and nuclear realms. They can create additional new non-nuclear threats to nuclear weapons and thereby generate strategic instability. Sauer is also concerned about the deployment of autonomous systems. He fears that their interaction on the battlefield could cause unwanted military escalation within a split second, which could have dramatic consequences in the case of a face-off between nuclear-armed states.

In his essay (chapter 11), Jean-Marc Rickli of the Geneva Centre for Security Policy (GCSP) emphasizes the impact that AI could have on nuclear-armed states' perceptions of each other's capabilities. He argues that advances in AI technology present the prospect of disruption in the realm of strategy as they undermine the confidence that states place in their second-strike capability. He points out that the trickiest and most deceitful destabilizing effect of AI lies in the fact that a nuclear-armed state could easily misperceive its adversary's capabilities and intention. The belief that the adversary could defeat the nuclear-armed state's second-strike capability might be sufficient for that state to take destabilizing action. Rickli thus argues that it should be a high priority for the nuclear powers to communicate clearly about their AI capabilities.

Justin Bronk of the Royal United Service Institute (RUSI) focuses in his essay (chapter 12) on the case of unmanned combat aerial vehicles (UCAVs). He discusses the factors that prompt nuclear-armed states and other major military powers to develop and acquire these systems. For Bronk, even if UCAVs are confined to conventional warfare missions, their proliferation could have an impact on strategic stability and increase the risk of a nuclear escalation. UCAVs could threaten the ground-based air defences that a number of nuclear-armed states use to defend their critical national assets. In a crisis situation, the fear of being attacked by an enemy's UCAVs could exert pressure on a nuclear-armed state that is not able to effectively defend itself with conventional means to use nuclear weapons. Another potentially destabilizing factor that Bronk highlights is the fact that UCAVs will, necessarily, be autonomous weapon systems. He points out that there are critical operational, legal and ethical questions associated with

the use of autonomy in combat missions. For Bronk, the member states of the North Atlantic Treaty Organization (NATO) and their allies should take the lead in the establishment of norms around the use of UCAVs and autonomous weapon systems more generally.

Unlike the other contributions, Shahar Avin of the University of Cambridge and S. M. Amadae of the University of Helsinki do not focus their essay (chapter 13) on the direct, first-order effect of AI on the traditional nuclear deterrence architecture. Rather, they draw attention to second- and higher-order effects, in particular the new vulnerabilities that could be introduced in peripherical systems that support nuclear command and control such as the vast computer systems that gather and analyse intelligence relevant to nuclear decision-making. Avin and Amadae also discuss how machine learning could be used to conduct cyber operations against nuclear weapon systems and influence campaigns on personnel working directly or indirectly with nuclear weapons and related systems as well as broader public opinion. They conclude by stressing that national and international measures to improve the cybersecurity of the trusted computing base of nuclear deterrence will be of paramount importance to reduce nuclear risk.

Finally (in chapter 14), Anja Kaspersen and Chris King of the UN Office for Disarmament Affairs (UNODA) shared their personal views on how states could work together and with partners, new and old, to deal with the challenges and opportunities that AI and emerging technologies generate for strategic stability and nuclear risks. They argue that one priority should be advancing states' understanding of how developments in technology could increase nuclear risk. They propose in that regard that technology-based risk should be taken into consideration in ongoing and future nuclear risk-reduction discussions. Kaspersen and King argue that the international community should work towards the development of politically binding transparency and confidence-building measures such as an agreement to not interfere with command-and-control structures. They believe at the same time that traditional arms-control measures should also be supported by 'soft' law or self-regulatory approaches to responsible innovation. Finally, Kaspersen and King stress that AI brings also opportunities for disarmament and non-proliferation. They explain for instance that AI could help to monitor for nuclear tests and to prevent illicit procurement of weapons of mass destruction.

The volume concludes (in chapter 15) with a summary of the key conclusions drawn from the essays. Notably, the chapter discusses the extent to which the contributors agree on the opportunities and risks that the AI renaissance brings to the field of nuclear weapons, nuclear doctrines and strategic stability in the Euro-Atlantic context.

Part I. Demystifying artificial intelligence and its military implications

This decade's renaissance of artificial intelligence (AI) has received growing attention from the popular media and policy community in recent years. However, despite the increasing number of publications and public events on the topic, widespread myths and misconceptions remain about what AI really is and what AI systems can do. This may seem surprising considering that this subfield of computer science has existed for more than half a century. The misconceptions are partly inherited from the way in which AI is depicted in popular culture, and in science fiction in particular. They are also reinforced by the way in which the media often tends to talk about technology: it uses eye-catching headlines that often exaggerate what the technology can do, for the best and for the worst. The problem with these misconceptions and misrepresentations is that they make it difficult—if not unproductive—to discuss the opportunities and risks of AI in general and in the security field in particular, as they obscure the true possibilities and limitations of AI technology. The following three essays thus aim to provide readers with a general but nuanced overview of what is currently happening in the field of AI.

The first essay (chapter 2) starts with a basic introduction to AI and discusses the components and implications of the current AI renaissance. Most notably, it explains why machine learning and autonomous systems are the focus of this volume. The subsequent essays provide the perspectives of a civilian engineer and a military engineer on the current state of AI technology, focusing on the case of autonomous systems. Dimitri Scheftelowitsch (in chapter 3) discusses in general terms what types of task state-of-the-art AI systems can and cannot undertake and why. Turning to the case of military systems, Martin Hagström (in chapter 4) puts the role of automation and autonomy in military systems in a historical perspective and assesses the role that machine learning could play in future military systems, including military autonomous systems.

VINCENT BOULANIN

2. Artificial intelligence: A primer

VINCENT BOULANIN

The concept of artificial intelligence (AI) was coined in the mid-1950s by John McCarthy, who defined it broadly as the 'science and engineering of making intelligent machines'.[1] However, the concept of AI means different things to different people, partly because its subject matter—intelligence—is hard to define.[2] Moreover, AI has, as one observer had put it, 'a bit of a Hollywood problem': popular culture has made it easy to talk about AI applications—from robots to smart digital assistant—but at the same time has 'skew[ed] expectations'.[3] There is a vast gap between the reality of what AI can do and the understanding, expectations and fears of the public, including often informed policymakers.

As a way to debunk potential misconception and misunderstanding around what AI is, this essay provides a basic introduction to this field of science and technology. It starts (in section I) by looking at how AI has been defined, how AI approaches have varied over time, and what is at stake with the current AI renaissance. It then presents two AI-related developments that arguably need particular attention when considering the potential of AI in the military sphere: machine learning (section II) and autonomous systems (section III).

I. What is AI?

The concept and its interpretations

For the majority of AI researchers, AI is about making machines capable of mimicking capabilities that are usually associated with human intelligence, such as observing the world through vision, processing natural language or learning. Some AI researchers differentiate between so-called narrow (or weak) AI and artificial general intelligence (AGI, or strong AI).

AGI is general-purpose AI: AI that would match—if not outperform—a human's ability to make sense of the world and to develop an understanding of its environment. It is the kind of AI that is typically depicted in popular culture in films such as *The Terminator*, *Blade Runner* or *2001: A Space Odyssey*. AGI has always fascinated AI researchers, but its design remains an unresolved technical challenge. There are, in fact, strong disagreements as to whether AGI will be ever

[1] Pearl, A., 'Homage to John McCarthy, the father of artificial intelligence (AI)', Artificial Solutions, 2 June 2017.

[2] Dale, R., 'An introduction to artificial intelligence', ed. A. M. Din, SIPRI, *Arms and Artificial Intelligence: Weapons and Arms Control Applications of Advanced Computing* (Oxford University Press: Oxford, 1987), p. 33.

[3] Madhavan, R., 'Understanding the societal impact of autonomous technologies', *IEEE Future Directions*, Nov. 2016.

be possible. Even the most optimistic AI researchers admit that AGI programs will remain in the realm of science fiction for the foreseeable future.[4]

Narrow AI has been around for decades and is the type of AI that is widely used today. Narrow AI systems are complex software programs that can execute discrete 'intelligent' tasks such as recognizing objects or people from images, translating language, or playing games. Narrow AI systems execute complex calculations, but they are brittle in nature—they are limited by the boundaries of their programming and they only work, at least reliably, for the intended tasks and operating environment. This is the type of AI that this volume refers to when it talks about AI.

For the purposes of this essay—and this volume more generally—AI is a catch-all term that refers to a wide set of computational techniques that allow computers and robots to solve complex, seemingly abstract problems that had previously yielded only to human cognition.[5] A key definitional element to bear in mind is that (narrow) AI is not a definite, unified technology. Instead, it is a portfolio technology that encompasses a wide variety of enabling applications which may be used to 'cognify' (i.e. give some form of cognitive capability to) multiple types of technology, including weapon technology.

AI is often depicted as a new or emerging technology, but it is not. As an academic discipline, it is more than half a century old. Narrow AI applications have been used for civilian and military purposes since the 1960s.[6]

A constant in the public debate has been the use of the concept of AI to refer to the newest computer technologies. Once these technologies have been widely deployed and adopted, they are no longer thought of or depicted as AI technologies. In other words, the frontier of AI is always moving: what is considered AI today may be considered normal software technology in the near future.

The current AI hype: machine learning and autonomous systems

Since the 1950s, the field of AI has gone through several 'hype cycles': each period of major success (an 'AI summer') was inevitably followed by a period of disillusions (an 'AI winter') as the new and promising approach of AI eventually failed to match its early expectations.[7] These AI winters typically resulted in funding cutbacks.

Since the beginning of this decade, the field of AI has been experiencing a new summer due to a breakthrough in machine learning. This approach to AI

[4] Ready, C., 'Kurzweil claim that singularity will happen by 2045', Futurism, 5 Oct. 2017.

[5] International Panel on the Regulation of Autonomous Weapons (IPRAW), *Focus on Computational Methods in the Context of LAWS*, 'Focus on' Report no. 2 (German Institute for International and Security Affairs: Berlin, Nov. 2017).

[6] Dale (note 2); Russel, S. and Norvig, P., *Artificial Intelligence: A Modern Approach*, 3rd edn (Pearson Education: Harlow, 2014); Kit, P., 'What should we learn from past AI forecasts?', Open Philanthropy Project, May 2016; and Armstrong, S. and Sotala, K., 'How we're predicting AI—or failing to', eds J. Romportl et al., *Beyond AI: Artificial Dreams*, Proceedings of the International Conference 'Beyond AI 2012', Pilsen, Czechia, 5–6 Nov. 2012 (University of West Bohemia: Pilsen, 2012), pp. 52–75.

[7] On hype cycles see Gartner, 'Gartner hype cycle', [n.d.]; and Kit (note 6).

software development has been around since the beginning of AI research but has greatly benefited in the past decade from the progress of computer power and the increasing availability of digital data.[8]

Like previous AI summers, success stories about what current AI systems can achieve have channelled major interest and investment towards the most promising approach to AI engineering—currently machine learning—but also toward concrete applications that could be derived from it. One area of application that has received significant attention in recent years is autonomy. Autonomy (or 'machine autonomy') can be defined as the ability of a machine to execute a task or tasks without human input, using interactions of computer programming with the environment.[9] An autonomous system is, by extension, usually understood as a system—whether hardware or software—that, once activated, can perform some tasks or functions on its own. Progress in machine learning has opened major opportunities for increasing autonomous capabilities within systems and developing commercially viable autonomous systems, such as AI voice assistants, autonomous cars or autonomous weapons.

II. Machine learning: A key enabler of the AI renaissance

What is machine learning?

Machine learning is an approach to software development that first builds systems that can learn and then teaches them what to do using a variety of methods (i.e. supervised learning, reinforcement learning or unsupervised learning).[10] It removes the need for hand-coded programming, whereby humans hard-code software features into the systems.[11]

Machine learning was a marginal subfield of AI in the 1960s and 1970s as it was of limited practical use. In the 1980s and 1990s the digitalization of many industries and the emergence of large data sets—on which machine learning systems can be trained—reignited interest and inspired the development of new techniques. Among these were refined versions of a method known as 'artificial neural networks', which draws on knowledge of the human brain, statistics and applied mathematics.

The main advantage of machine learning compared to traditional hand-coded programming is that a human does not have to explicitly define the problem to be solved by the software and the way in which it solves the problem. Hand-coded programming usually requires a great deal of research on how the world works. In order to develop the model and rules that will govern the behaviour of an

[8] Knight, W., 'There is a big problem with AI', *MIT Technology Review*, 11 Apr. 2017.

[9] This definition is based on one previously proposed by Andrew Williams. Williams, A., 'Defining autonomy in systems: challenges and solutions', eds A. P. Williams and P. D. Scharre, *Autonomous Systems: Issues for Defence Policymakers* (NATO Headquarters Supreme Allied Commander Transformation: Norfolk, VA, 2015), pp. 27–62.

[10] For more details see Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017).

[11] Knight (note 8).

Box 2.1. Deep learning

Deep learning is an approach to machine learning whereby the system 'learns' how to undertake a task in supervised, semi-supervised or unsupervised ways. It transforms raw data input into abstract representations (features) that can be effectively exploited in machine learning tasks, such as recognizing an object in an image.

The strength of deep learning is in its ability to introduce representations that are expressed in terms of other, simpler representations. In other words, it allows the computer to build complex concepts from simpler concepts. A deep-learning system can, for instance, represent the concept of an image of a person by combining simple concepts, such as corners and contours.[a]

The success of deep learning was supported by two trends. First was the widespread commercialization of graphics processing units (GPUs), a type of computer chip that is well suited for machine learning operations. Second, and perhaps more importantly, was the widespread democratization of the Internet and social media, which led to an explosion in the volumes of the digital data on which machine learning algorithms can be trained.

In recent years, deep learning has become the most fashionable approach to artificial intelligence (AI) engineering, but this does not mean that it has totally supplanted other approaches. Many cutting-edge AI applications do not use deep learning—or even machine learning. Many continue to rely on relatively old-fashioned hard-coded expert knowledge, while others may use established machine learning methods such as Bayesian statistics or evolutionary algorithms.[b] Traditional AI programming methods will, in other words, continue to be relevant. In fact, some argue that truly intelligent machines can only be develop by combining machine learning with traditional programming that can introduce abstract knowledge and a layer of common sense reasoning.[c]

[a] Goodfellow, I., Bengio, Y. and Courville, A., *Deep Learning* (MIT Press: Cambridge, MA, 2016), p. 8.

[b] Hao, K., 'We analyzed 16,225 papers to figure out where AI is headed next', *MIT Technology Review*, 25 Jan. 2019.

[c] Thompson, C., 'How to teach artificial intelligence some common sense', *Wired*, 13 Nov. 2018.

autonomous system, the developing engineers often cooperate with scientists from other scientific fields, notably the natural sciences (e.g. neurosciences and physics) and the social sciences (e.g. psychology, linguistics and sociology). Hand-coded programming gets difficult for tasks and operating environments too complex for a human to model completely.[12] One such problem is image recognition. It is hard for human to express in mathematical terms what distinguishes pictures of cats from pictures of dogs. Machine learning allows engineers to bypass this difficulty, as it allows an image-recognition system to generate its own way of perceiving the differences.[13]

When used in non-technical context, the term 'learning' can sometimes be a source of confusion, as it invites an anthropomorphic interpretation. However, the way in which machine learning works has nothing to do with the way humans learn: machines learn by finding statistical relationships in past data.[14] Engineers use the term 'learning' for practical reasons: it is a concise and memorable way of describing a complex computing process.

[12] Kester, L., 'Mapping autonomy', Presentation at the CCW Informal Meeting of Experts on Lethal Autonomous Weapon Systems, Geneva, 11–15 Apr. 2016.

[13] Knight (note 8).

[14] Boulanin and Verbruggen (note 10).

The rise of deep learning as the dominant approach to AI engineering

In 2006 Geoffrey Hinton, a cognitive psychologist and computer scientist at the University of Toronto, co-authored an academic paper on what was deemed at the time to be a hopelessly musty academic problem in the AI community: how to make neural networks work more quickly and more effectively.[15] A neural network is a possible architecture for machine learning that was briefly popular in the 1980s but then became marginal and unfunded: reportedly, by the early 2000s, no more than five researchers specialized in neural networks.[16] Hinton was one of the last scholars in the AI community who believed that machine learning, and neural networks in particular, was a promising approach for the development of truly proficient AI systems. The paper demonstrated that, when combined in different layers, neural networks could be very powerful. When they published it, Hinton and his co-authors did not realize that they were onto something that was poised to not only transform the field of AI from the inside but also to reignite massive interest in machine learning and the field of AI more generally.[17]

The technique they came up with—later rebranded 'deep learning'—was to outperform traditional AI programming techniques by a wide margin (see box 2.1). AI systems that would rely on deep learning systematically beat existing AI systems at implementing tasks such as recognizing images or speech, translating language, or playing games such as chess.[18]

Industry and governments alike soon identified major opportunities and started to invest in deep learning and machine learning more broadly. Large technology companies such as Google and Facebook were the fastest to react. At the beginning of this decade, they bought the most innovative start-up companies that worked with deep learning.[19] They also recruited the few scholars that had remained active in this field and created dedicated research teams.[20] Many applications that these companies commercialize today—from smartphones to social media platforms—are powered by deep learning.

Governments also gradually developed a major interest in deep learning and the promise that it held in terms of enabling the development and use of practical and powerful applications in both the civilian and military spheres. Governmental research funding institutions, such as the United States' Defense Advanced Research Projects Agency (DARPA), started pouring money into machine learning-related research and development projects, while governments started

[15] Hinton, G. E., Osindero, S. and Teh, Y., 'A fast learning algorithm for deep belief nets', *Neural Computation*, vol. 18, no. 7 (July 2006), pp. 1527–54. See also Kurenkov, A., 'A "brief" history of neural nets and deep learning, part 4', *Medium*, 17 Feb. 2017.

[16] Allen, K., 'How a Toronto professor's research revolutionalized artificial intelligence', *Toronto Star*, 17 Apr. 2015.

[17] Somers, J., 'Is AI riding a one-trick pony', *MIT Technology Review*, 29 Sep. 2017.

[18] Gershgorn, D., 'The data that transformed AI research—and possibility the world', *Quartz*, 26 July 2017.

[19] E.g. Shu, C., 'Google acquires AI startup DeepMind for more than $500m', TechCrunch, 26 Jan. 2014.

[20] Malingo, T., 'The misguided rush of the academic AI brain drain', New Stack, 23 Aug. 2018.

Box 2.2. Generative adversarial networks

The generative adversarial network (GAN) is a new approach (invented in 2014) that involves two artificial neural network systems that can spar with each other to create ultra-realistic original image, audio or video content—something that machines have never been able to do properly before.

The two networks are trained on the same data set. One, known as the generator, is tasked with creating variations on images that it has already seen. The second, known as the discriminator, is asked to identify whether the example it sees is like the images on which it has been trained or a fake produced by the generator. Over time, the generator becomes so good that the discriminator cannot spot fakes.

The potential consequences of this breakthrough are both positive and negative. On the one hand, it could help a machine learning system to generate new data to train itself; on the other hand, it could create digital fakery for criminal or information warfare purposes.

Source: Condliffe, J., 'Duelling neural networks: by playing cat-and-mouse games with data, a pair of AI systems can acquire an imagination', *MIT Technology Review*, vol. 121, no. 2 (Mar./Apr. 2018).

internal discussions on what their strategy towards deep learning should be.[21] Between 2017 and early 2019 at least 17 countries released a national strategy or made a strategic policy announcement on AI.[22]

Opportunities and challenges

Whether from the industry or the governmental side, there is a common understanding that the AI renaissance that Hinton and his peers initiated has brought its share of opportunities and challenges.[23]

On the opportunity side, machine learning no longer needs to demonstrate its potential. It has already dramatically improved the ability of computers and robots to perceive the world, which has accelerated the development of autonomous systems such as self-driving cars and voice assistants.[24] Machine learning has also shown itself to be a powerful tool for data management. It can be used not only to classify data but also to find correlations in data that can then be used to make statistical predictions about future behaviour. Internet service providers, such as Google, Facebook or Baidu, routinely use machine learning to label and organize content such as text, images and videos, and to predict customer preferences.[25] National militaries are now trying to develop a similar capability to process

[21] US Defense Advanced Research Project Agency (DARPA), 'DARPA announce $2 billion campaign to develop next wave of AI technologies', 7 Sep. 2018.

[22] These 17 countries are Canada (Mar. 2017), Japan (Mar. 2017), Singapore (May 2017), China (July 2017), the United Arab Emirates (Oct. 2017), Finland (Dec. 2017), Denmark (Jan. 2018), Italy (Mar. 2018), France (Mar. 2018), the UK (Apr. 2018), South Korea (May 2018); Sweden (May 2018), India (June 2018), Mexico (June 2018), Germany (Nov. 2018), the European Union (Dec. 2018) and the USA (Feb. 2019). Dutton, T., 'An overview of national AI strategies', Medium, 18 June 2018.

[23] As identified by Scharre, P. and Horowitz, M. C., *Artificial Intelligence: What Every Policymaker Needs to Know* (Center for New American Security: Washington DC, June 2018), p. 10.

[24] Gershgorn, D., 'See the difference one year makes in artificial intelligence research', *Popular Science*, 31 May 2016.

[25] Marr, B., 'The amazing way Google uses deep learning AI', *Forbes*, 8 Aug. 2017.

intelligence data.[26] The ability of machine learning to identify patterns can also be used to detect anomalies in data. Cybersecurity companies increasingly rely on machine learning to detect new malware, while the surveillance industry is exploring the possibility of using machine learning to detect abnormal behaviour in CCTV footage.[27] Machine learning has also turned out to be useful for optimizing the performance of complex tasks, such as controlling large numbers of robots.[28] It can even generate new content such as original, ultra-realistic images, sounds or written stories. This has been achieved with a new machine learning-based approach known as a generative adversarial network (GAN, see box 2.2), which could reportedly become the next big thing in the AI community.[29] Machine learning holds great promise, but it also has significant shortcomings. The first—and perhaps most salient—relates to its dependence on data. Systems that are powered by machine learning are only as good as the data on which they are trained.[30] To be taught, a machine learning system needs to be provided with large volumes of real word examples (training data) and rules about the data relationships. In order to recognize a type of object in an image (e.g. a car, a bus or a dog), a computer vision system may need to be trained with millions of pictures of that type of object. The quality of the data on which the systems are trained is equally important. If the training data set is not representative, then the system might fail, might perform poorly, or might misinform human decisions and actions by reinforcing existing human biases or creating new ones.[31] Research has shown, for instance, that facial recognition systems trained with data sets that primarily include images of light-skinned men are more likely to misidentify faces of women or people with darker skin.[32] Companies and governments willing to develop and use machine learning therefore have a double challenge: they have to both find a sufficiently large set of training data and a way to ensure that this data is reliable. A number of recent news reports indicates that companies and governments alike are struggling with that challenge.[33]

Another fundamental, and related, problem is the fact the machine learning algorithm are opaque. Traditional handcrafted AI systems reason according to rules and logic, making the inner workings transparent to anyone who cares to

[26] One such system—the USA's Project Maven—is discussed in chapters 5, 6, 10 and 11 in this volume.

[27] Polyakov, A., 'Machine learning for cybersecurity 101', Towards Data Science, 4 Oct. 2018.

[28] Hüttenrauch, M., 'Guided deep reinforcement learning for robot swarms', Master's thesis, Technische Universität Darmstadt, Aug. 2016.

[29] Condliffe, J., 'Duelling neural networks: by playing cat-and-mouse games with data, a pair of AI systems can acquire an imagination', *MIT Technology Review*, vol. 121, no. 2 (Mar./Apr. 2018).

[30] Gershgorn (note 18); and Hao, K., 'We analyzed 16,225 papers to figure out where AI is headed next', *MIT Technology Review*, 25 Jan. 2019.

[31] Knight, W. and Hao, K., 'Never mind killer robots—here are six real AI dangers to watch out for in 2019', *MIT Technology Review*, 7 Jan. 2019; and Hao, K., 'This is how AI bias really happens—and why it's so hard to fix', *MIT Technology Review*, 4 Feb. 2019.

[32] Lohr, S., 'Facial recognition is accurate, if you're a white guy', *New York Times*, 9 Feb. 2018.

[33] See e.g. Buranyi, S., 'Rise of the racist robots—how AI is learning all our worst impulses', *The Guardian*, 8 Aug. 2017; Hao, K., 'Police across the US are training crime-predicting AIs on falsified data', *MIT Technology Review*, 13 Feb. 2019; and Hao, K., 'AI is sending people to jail—and getting it wrong', *MIT Technology Review*, 21 Jan. 2019.

examine the code. In contrast, a machine learning system, particularly one that relies on deep neural networks, operates like a black box.[34] The input and the output of such a system are observable, but the computational process leading from one to the other is difficult for humans to understand. It is particularly difficult for humans to understand what such a system has learned and hence how it might react to input data that is different from that used during the training phase.[35] The lack of transparency and explainability of these systems in turn creates a fundamental problem of predictability. A machine learning system might fail in ways that were unthinkable to humans because the engineers do not have a full understanding of its inner working. In the context of weapon systems, this unpredictability could have dramatic consequences. The lack of transparency is also problematic from a regulatory standpoint as it makes complex the task of identifying the source of a problem and attributing responsibility when something goes wrong.[36]

Moreover, AI systems trained with machine learning may outperform human for many tasks, but they still lack what humans understand as basic common sense. Computer vision systems, for instance, do not perceive a pattern at an abstract level, like a human would. They just see a correlation between a group of pixels. A facial recognition system could not tell the difference between an actual person and a picture of a picture within an image: in both cases it would be a positive identification. One study also recently demonstrated that variations in an image that are imperceptible to the human eye could cause an image-recognition system to completely mislabel the object or people in the image (e.g. mistaking a lion for a car or a building).[37] Another study has demonstrated that it is easy to produce images that are completely unrecognizable to humans but that computer vision software believes to be a recognizable object with over 99 per cent confidence.[38] In other words, machine learning systems may not be reliable: they can easily be fooled (which is particularly problematic if used in an adversarial context such as an armed conflict) or they may fail in unpredictable ways according to human standards.

In sum, while recent advances in machine learning have created important opportunities for the development of highly efficient AI systems, including autonomous systems, machine learning is still, in several regards, an immature technology. There remain many technical practical challenges associated with the use of machine learning methods. The fundamental question that developers, users and regulators are currently struggling with is how to ensure the responsible

[34] Knight, W., 'The dark secret at the heard of AI', *MIT Technology Review*, 11 Apr. 2017.

[35] Righetti, L., 'Emerging technology and future autonomous systems', International Committee of the Red Cross (ICRC), *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, Expert meeting, Versoix, Switzerland, 15–16 Mar. 2016 (ICRC: Geneva, Aug. 2016), pp. 36–39.

[36] Tobey, D., 'Explainability: where AI and liability meet', DLA Piper, 25 Feb. 2019.

[37] Szegedy, C. et al., 'Intriguing properties of neural networks', arXiv, 1312.6199, version 4, 19 Feb. 2014.

[38] Nguyen, A., Yosinski, J. and Clune J., 'Deep neural networks are easily fooled: high confidence predictions for unrecognizable images', *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Proceedings, 7–12 June 2015 (Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, 2015), pp. 427–36.

adoption and use of this technology. This question is particularly pressing for systems and applications that are safety critical (e.g. cars, aeroplanes and weapons) or that could have a societal impact (e.g. health and educational services).

III. Autonomy: A key by-product of the AI renaissance

What is autonomy?

As a technology area, autonomy is related to but distinct from AI. While machine learning can be depicted as the key ingredient of the current renaissance of AI (and the associated hype), autonomy can be portrayed as one of its key by-products. Autonomous systems ranging from AI assistants (e.g. Amazon's Alexa or Apple's Siri), via self-driving cars and auto-piloted unmanned aerial vehicles (UAVs) to autonomous weapons are among the most debated technological developments deriving from the current AI renaissance and with the highest level of media attention.[39]

As discussed above, 'autonomy' can be understood as the ability of a machine to execute a task, or tasks, without human input, using interactions of sensors, computer programming and actuators with the environment.[40] However, the concept of autonomy and autonomous systems more generally mean different things to different people, primarily because autonomy is a relative notion that can be interpreted in several ways. The level of autonomy of a system can be analysed from three different and independent perspectives (see figure 2.1): (*a*) based on the extent to which humans are involved in the execution of the task carried out by the system; (*b*) based on the extent to which the system can exercise control over its own behaviour and deal with uncertainties in its operating environment; and (*c*) based on the number and types of functions that are automated.[41]

This definitional uncertainty appears clearly in this volume: the contributors sometimes use different terminology and metrics to talk about autonomy. For example, some experts make a distinction between automatic, automated and autonomous systems, while others use these terms interchangeably (see box 2.3). For the analytical purposes of this volume, it is not necessary to resolve that conceptual debate: the level of autonomy of systems can be analysed through different lenses and each lens has its own analytical value.

[39] Hao, K., 'One day your voice will control all your gadgets, and they will control you', *MIT Technology Review*, 11 Jan. 2019; 'Autonomous weapon and the new laws of war', *The Economist*, 17 Jan. 2019; and Salesky, B., 'A decade after DARPA: our view on the state of the art in self-driving cars', *Medium*, 16 Oct. 2017.

[40] Williams (note 9), pp. 55–56.

[41] As identified by Scharre, P., 'The opportunity and challenge of autonomous systems', eds Williams and Scharre (note 9), pp. 3–26. See also Thrun, S., 'Toward a framework for human-robot interaction', *Human–Computer Interaction*, vol. 19, nos 1–2 (June 2004), pp. 9–24; and Boulanin and Verbruggen (note 10).

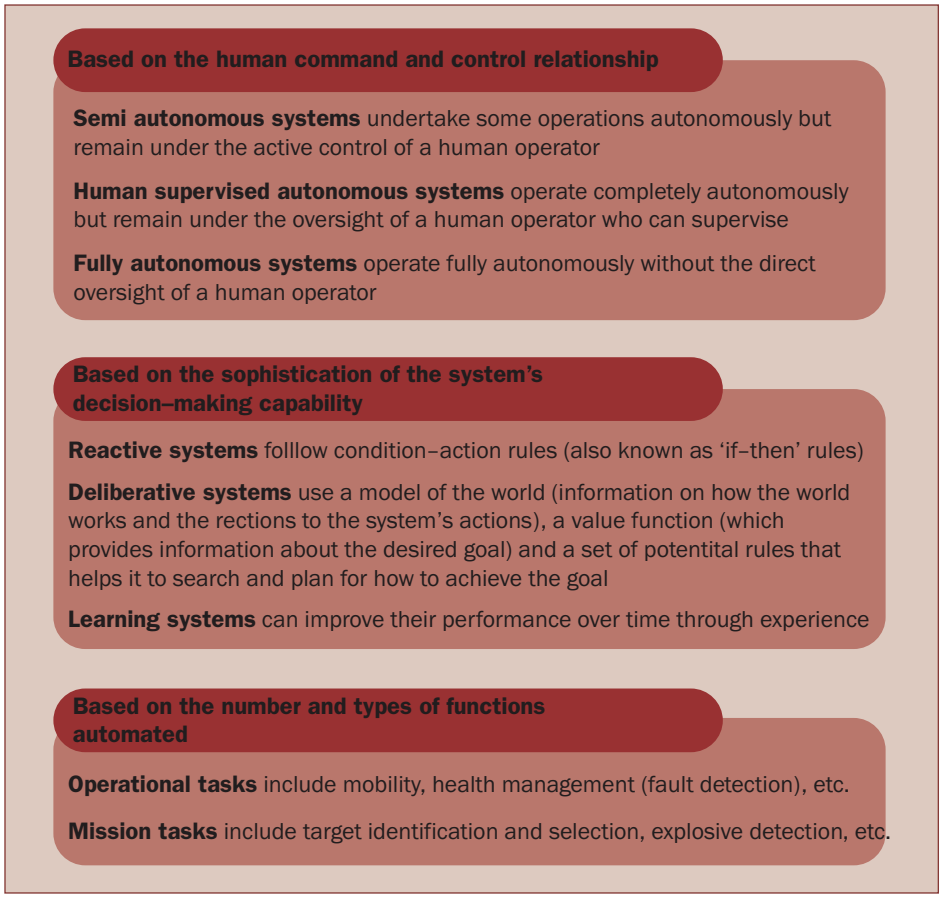

Figure 2.1. Approaches to the definition and categorization of autonomous systems

Source: Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017).

Opportunities and challenges[42]

Advances in autonomy are generating great expectations in both the civilian and military spheres as they enhance the usefulness and reliability of robotics systems, which in turn could generate significant economic and operational benefits. Companies, governmental institutions and the military alike could achieve greater manpower efficiency by increasing their reliance on robotic systems.[43] Advances in autonomy could also allow them to overcome a number of operational challenges associated with manned operations or the use of teleoperated systems.[44]

[42] This section is based on Boulanin and Verbruggen (note 10), pp. 61–82.

[43] US Department of Defense (DOD), Defense Science Board, *The Role of Autonomy in DoD Systems*, Task Force Report (DOD: Washington, DC, July 2012); and Scharre, P., *Robotics on the Battlefield*, part II, *The Coming Swarm* (Center for New American Security: Washington, DC, Oct. 2014).

[44] eds Williams and Scharre (note 9).

Box 2.3. Automatic, automated, autonomous

Automatic

The label 'automatic' is usually reserved for systems that mechanically respond to sensory input and step through predefined procedures, and whose functioning cannot accommodate uncertainties in the operating environment. An example of this is a robotic arm used in the manufacturing industry.

Automated versus autonomous

Machines that can cope with variations in their environment and exercise control over their actions can be described as either automated or autonomous. What distinguishes an automated system from an autonomous system is a contentious issue.

Some experts see the difference in terms of the degree of self-governance. They view autonomous systems merely as more complex and intelligent forms of automated system.

Others see value in making a clear distinction between the two concepts. A report from the US Defense Science Board presents an automated system as a system that is governed by 'prescriptive rules that permit no deviations'.[a] This means that the system logically follows a pre-defined set of rules in order to provide an outcome; its output is predictable if the set of rules under which it operates is known. In contrast, an autonomous system is able to 'independently compose and select among different courses of action to accomplish goals based on its knowledge and understanding of the world, itself, and the situation'.[b]

While the distinction between the terms automatic, automated and autonomous can be conceptually useful, in practice it is difficult to determine which of the three categories a system belongs to. Moreover, the definitions of and boundaries between these three categories remain contested within and between the expert communities.

[a] US Department of Defense (DOD), Defense Science Board, *Report of the Defense Science Board Summer Study on Autonomy* (DOD: Washington, DC, June 2016), p. 4.

[b] US Department of Defense (note a), p. 4.

Source: Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017).

Autonomy means that a robotic system can execute some tasks much faster than any human or human-operated robot ever could, which for the military is particularly attractive for time-critical missions or tasks such as air defence, air-to-air combat or cyber-defence. Autonomy can make robotic systems far more agile from a command-and-control perspective and reduce the need to maintain a constant communications link between the robot and the military command. It can also allow the military to reduce the number of human operators and analysts needed to oversee the system and process information. Autonomy is also useful for so-called dull, dirty and dangerous (3D) missions as it removes limitations (e.g. fatigue, boredom, hunger or fear) that may make human performance deteriorate over time. Autonomy also gives systems greater reach. It grants access to operational theatres that were previously inaccessible to remote-controlled systems or too risky for manned operations. These include areas protected by anti-access/area-denial (A2/AD) systems and areas with harsh operating environments for humans (and where communication is limited), such as deep water, the Arctic and, potentially, outer space. Finally, autonomy also provides new opportunities for collaborative operations as it permits weapon systems to

operate in large groups, or 'swarms', in a much more coordinated, structured and strategic way than if they were individually controlled by a human operator.

The advances in autonomy are, however, raising a wide spectrum of ethical, legal and security concerns, which apply to both civilian and military applications.[45] From an ethical standpoint, the development of autonomy in safety-critical systems such as cars or weapons raises the vexed question of whether, and to what extent, autonomous systems should be trusted to operate outside of direct human control and supervision. The ongoing intergovernmental discussion on lethal autonomous weapon systems (LAWS) in the framework of the 1980 Convention on Certain Conventional Weapons (CCW Convention) and the debate around car accidents involving semi-autonomous cars have shown that there is no simple answer to that question.[46] The question of the balance between autonomy and human control also has profound and complex legal implications, particularly with regard to the attribution of individual criminal responsibility: who should be prosecuted when a self-driving car or autonomous weapon causes harm? Some legal experts argue that the proliferation of autonomous systems is taking place in a legal vacuum or grey zone, which could make the determination of legal accountability difficult in case of a deadly incident.[47] Advances in autonomy also create new security risks. In addition to the increasing vulnerability to cyberattacks, the limitations of existing autonomous systems in terms of perceptual and decision-making intelligence could easily be exploited by a malevolent actor who could defeat a system by simply spoofing the sensors or control systems.[48]

More broadly, the increasing adoption of and reliance on autonomous systems is bound to ignite profound societal changes.[49] Among other effects, it will change the way that companies, governmental agencies and the military operate. Taking the case of an air force as an example, replacing manned combat aircraft with autonomous unmanned aerial systems will necessitate a change in the way that personnel are selected, trained and meant to operate—with control moving from a pilot to a remote operator and then to a systems supervisor—which in turn could cause major changes in professional culture.[50]

[45] Cath, C. et al., 'Governing artificial intelligence: ethical, legal and technical opportunities and challenges', *Philosophical Transactions of the Royal Society A*, vol. 376, no. 2133 (Nov. 2018).

[46] Boulanin, V., 'Mapping the debate on LAWS at the CCW: taking stock and moving forward', EU Non-proliferation Paper no. 49, EU Non-proliferation Consortium, Mar. 2016; Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects (CCW Convention), opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983; and Bhuyian, J., 'Uber's semi-autonomous cars detected the pedestrian six seconds before the fatal crash, a federal agency says', Recode, 25 May 2018.

[47] Docherty, B., *Mind the Gap: The Lack of Accountability for Killer Robots* (Human Rights Watch: New York, 2015).

[48] Versprille, A., 'Army still determining the best use for driverless vehicles', *National Defense*, June 2015; and Endsley, M. R., *Autonomous Horizons: System Autonomy in the Air Force—A Path to the Future*, vol. 1, *Human–Autonomy Teaming* (US Air Force, Office of the Chief Scientist: Washington, DC, 2015), p. 5.

[49] Wright, N., 'Three distinct AI challenges for the UN', AI & Global Governance, United Nations University, Centre for Policy Research, 7 Dec. 2018.

[50] For detailed discussion see Boulanin and Verbruggen (note 10), pp. 69–73; and chapter 12 in this volume.

IV. Conclusions

In a June 2018 essay in *The Atlantic*, Henry Kissinger, a former US Secretary of State, argues that 'The Enlightenment started with essentially philosophical insights spread by a new technology. Our period is moving in the opposite direction. It has generated a potentially dominating technology in search of a guiding philosophy.'[51] The dominating technology that he refers to is artificial intelligence. Recent advances in machine learning have unlocked numerous possibilities, including that of creating increasingly autonomous systems. However, developers and users alike have only begun to work out how to make the best use of it, and also how they should not use it.[52]

Big AI companies have come together and launched several initiatives that are meant to lead the global reflection on these questions, notably the Partnership on AI and OpenAI.[53] The Institute of Electrical and Electronics Engineers (IEEE), the world's largest association of engineers, has also started a conversation, which is not limited to its members, on the responsible development of AI systems.[54] Some states, through a national strategy on AI, are trying to work out a national approach to these issues, and also to identify concrete solutions. For example, the French national strategy on AI calls for funding of research and development work that would make machine learning systems more explainable in order to make them more socially acceptable.[55] It also supports the inclusion of ethics in training for AI engineers and researchers and the implementation of social impact assessments on new machine learning systems.

So far, most of these efforts have been connected to the civilian development of AI technology. The debate on what should be the norms and principles for responsible development of AI systems in the military sphere has gained momentum but has been constrained by the framework of the ongoing CCW debate on LAWS, which, of course, focuses on conventional weapons. The remainder of this volume opens up that conversation to the field of nuclear weapons and doctrines. The essays that follow are intended to lay the foundations for a constructive conversation on the possibilities and challenges that the current AI renaissance will bring to the realm of nuclear weapons and doctrines and to help identify possible norms and principles for the responsible development of AI systems in that context.

[51] Kissinger, H., 'How the enlightenment ends', *The Atlantic*, June 2018.

[52] Vogt, H., 'Artificial intelligence rules more of your life. Who rules AI?', *Wall Street Journal*, 13 Mar. 2018; and Hao, K., 'Why AI is a threat to democracy—and what we can do to stop it', *MIT Technology Review*, 26 Feb. 2019.

[53] Partnership on AI, 'About us', [n.d.].

[54] Madhavan (note 3); IEEE Standards Association, 'The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems', [n.d.]; and IEEE Global Initiative, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, version 2 (Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, [Dec. 2017]).

[55] Villani, C., *For a Meaningful Artificial Intelligence: Toward a French and European Strategy*, (Conseil national du numérique: Paris, Mar. 2018).

3. The state of artificial intelligence: An engineer's perspective on autonomous systems

DIMITRI SCHEFTELOWITSCH

In recent years, the topic of autonomous (robotic) systems has moved from an area of mostly academic research to an issue of public interest. Progress in the design and development of high-performance computing architectures in compact form factors (most notably modern graphics processing units) has enabled the design and development of autonomous devices with versatile application contexts. These have now gone beyond proof-of-concept and design studies. Think tanks focused on existential risks (e.g. the University of Cambridge Centre for the Study of Existential Risk) have also turned their attention to potential risks arising from the use of autonomous systems.

This essay provides an introduction to the topic of autonomy (section I) and its current and potential applications (section II). It then gives an overview of the technical and security challenges related to the design and use of autonomous systems, especially in the military (and, specifically, nuclear) domain (section III).

I. Autonomy: A primer

In order to discuss issues of autonomy, it is helpful to define what an autonomous system is: it is a system with the capability to observe its environment, plan a sequence of actions based on those observations and some previously acquired model of the world, and then execute the computed sequence with little or no interaction with a human operator.[1]

This implies that an autonomous system has an input data stream from sensing equipment, which is then processed into an internal representation of a problem domain (called a state), a model of possible consequences of its own actions, and a scheduling algorithm that computes an order of actions that fulfils an initially given goal. In virtually all applications, the goal has to be given implicitly or explicitly by the system's designer or operator. The goal itself is a function that maps the observed state to either a qualitative judgement of whether the current state fulfils some given property (i.e. the system judges whether the goal has been achieved) or a quantitative value that reflects a utility value of a state (i.e. the system measures the amount of energy, data or some purely virtual currency that it has accrued).

It is important to note that the distinction between autonomy and the similar notion of automation is fuzzy at best. Automation is often defined as the capability

[1] This definition follows Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017), pp. 7–12. On the definition of autonomy see also chapter 2 in this volume.

of a system to perform simple, easy-to-define tasks while autonomy implies at least some capacity to observe, plan and make decisions. However, this distinction is not as clear-cut as the wording may imply, as the class of tasks that are 'easy' to define and automate becomes larger (not least because of progress in technology). However, in what follows, this distinction is not crucial.

II. Applications

Autonomous devices can be used in several current and future domains. The most prominent are, naturally, robotic applications such as autonomous tour guides.[2] However, there are also actual and potential non-robot, stationary applications.

Robotic applications

The best-researched aspect of autonomous robotic devices are self-driving robots. For virtually all aspects of transportation, there exists at least research on autonomous navigation; in some cases, systems are available commercially. Self-driving cars have been most prominent, at least since the Grand Challenges held by the US Defense Advanced Research Projects Agency (DARPA) in 2004–2005 and its Urban Challenge in 2007.[3] However, unmanned sailing, underwater and aerial vehicles also currently have the capability to navigate in their respective domain, which enables advanced applications such as autonomous farming robots.[4]

Two further scenarios where partial or full autonomy is achievable are industrial robotics in the broader sense and surgical robotics. Surgical robotics is a specialization with pronounced decision-making capabilities. A current topic of research is accurate recognition of tissue to be removed in non-invasive surgery.[5] Robotic surgery that requires machine-precision cutting is also being developed.[6] For industrial robotics, the above-mentioned farming applications integrate robotics with autonomous navigation.

In the consumer market, personal robots are a particular application of robotics in human–robot communication that integrate sensing and decision-making in a

[2] Burgard, W. et al., 'The interactive museum tour-guide robot', *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98 (American Association for Artificial Intelligence: Menlo Park, CA, 1998), pp. 11–18.

[3] US Defense Advanced Research Projects Agency (DARPA), 'The Grand Challenge', [n.d.].

[4] Saildrone; Miller, P. A. et al., 'Autonomous underwater vehicle navigation', *IEEE Journal of Oceanic Engineering*, vol. 35, no 3 (July 2010), pp. 663–78; and López, J. et al., 'Comparative study of autonomous aerial navigation methods oriented to environmental monitoring', eds J. C. Mendes Carvalho et al., *Multibody Mechatronic Systems* (Springer: Cham, 2018), pp. 305–14; and Flourish, 'Flourish project', [n.d.].

[5] Alpers, J. et al., 'CT-based navigation guidance for liver tumor ablation', eds. S. Bruckner et al., *Eurographics Workshop on Visual Computing for Biology and Medicine* (Eurographics Association: Goslar, 2017).

[6] Shademan, A. et al., 'Supervised autonomous robotic soft tissue surgery', *Science Translational Medicine*, vol. 8, no. 337 (4 May 2016), research article no. 64.

robotic cyber–physical system for human interaction tasks. The impact on society has yet to be fully researched.[7]

Stationary applications

The autonomous applications described above are cyber–physical systems in the sense that they imply a robot interacting with its environment. However, autonomy does not necessarily imply a robotic system. Delegation of decision-making to a computer also occurs in non-robotic, digital applications such as automatic trading and disaster detection and warning.[8]

In a military context, a similar task is performed by the early-warning radar installations that monitor incoming missiles and issue alerts to control staff. Future applications include air traffic control, that is, autonomous air traffic routing and communication to pilots, which can also be used in the military domain for automation of air and missile defence.[9]

III. Challenges

Despite the recent successes, the design of an autonomous system that can be used in practice is a considerable engineering, mathematical and political challenge. The reasons for this lie not necessarily in the autonomous decision-making as such, since it is often easy to provide an appropriate mathematical model, but in the various other, not necessarily technical, aspects of autonomy. The following subsections identify important issues that need to be solved in order to design a dependable autonomous system (i.e. a system that performs accurately in a reliable and safe fashion), as well as policy questions that have to be answered at the design stage when discussing safety requirements, usage scenarios and the goals of the system.

Observation and interpretation

The observation task requires the autonomous system to accurately interpret sensory data in order to estimate the current state of the world and its evolution. In complex interactions with large domains, such as autonomous driving, this might prove difficult, because dramatically different situations can appear to differ in only marginal aspects. For example, the German theoretical driving test

[7] See e.g. Kory Westlund, J. M. et al., 'Measuring young children's long-term relationships with social robots', *Proceedings of the 17th ACM Conference on Interaction Design and Children*, IDC '18 (Association for Computing Machinery (ACM): New York, 2018), pp. 207–18.

[8] Huang, B. et al., 'Automated trading systems statistical and machine learning methods and hardware implementation: a survey', *Enterprise Information Systems*, vol. 13, no. 1 (2019), pp. 132–44; and Boukerche, A. and Coutinho, R. W. L., 'Smart disaster detection and response system for smart cities',*2018 IEEE Symposium on Computers and Communications*, ISCC 2018 (Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, 2018), pp. 1102–107.

[9] Mahboubi, Z. and Kochenderfer, M. J., 'Continous [sic] time autonomous air traffic control for non-towered airports', *2015 54th IEEE Conference on Decision and Control (CDC)* (Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, 2015), pp. 3433–38.

includes an image of a residential road with a ball on it. The ball is a minor feature of the perceived picture, yet it signals to the driver that the situation has become potentially dangerous, since a child may appear on the road at any moment.

Acting under uncertainty and interference

The theoretical view of the world is often a simplification of physical reality in terms of mathematical models. While this is well known, it may make the planning problem particularly hard. Two concrete problems arise.

The first problem is uncertainty in the model: by modelling the physical domain in mathematical formulas, the engineer has to rely on estimated parameters, which may vary with time, be insufficiently precise or provide incomplete information. The domain thus becomes partially observable (i.e. the state is not completely visible to the system's sensors) and requires different methods to tackle the planning problem.[10]

The second problem also stems from a different aspect of incomplete information. Autonomous systems typically act in dynamic environments with other agents. These other agents, in turn, may pursue similar or different goals and may or may not cooperate, which is not necessarily known beforehand. The existence of more agents with their own goals may alter the mechanics of the environment significantly and lead to potentially catastrophic outcomes. This scenario is not just hypothetical: a major example is flash crashes on trading markets, where the actions of trading algorithms provoke a self-sustaining selling feedback loop.[11] More generally, similar dynamics can appear in situations with several independent autonomous agents with non-cooperative goals. These must be dealt with by expanding the model accordingly, which requires far more computational resources.[12]

Goal specification

The last and most important issue comes from a non-computational context. Even given perfect observation, perfect situation awareness, perfect modelling and perfect computational decision-making capabilities, the actions of an autonomous system are defined by its goal statement. Although this is a trivial observation, it needs to be realized that, first, a goal must be formulated in machine-readable terms and, second, a computer system will follow exactly the goals that have been defined by its operator, no more and no less. In other words, the operator must be

[10] Kaelbling, L. P., Littman, M. L. and Cassandra, A. R., 'Planning and acting in partially observable stochastic domains', *Artificial Intelligence*, vol. 101, no. 1–2 (May 1998), pp. 99–134.

[11] Kirilenko, A. et al., 'The flash crash: high-frequency trading in an electronic market', *Journal of Finance*, vol. 72, no. 3 (June 2017), pp. 967–98.

[12] Coulombe, M. J. and Lynch, J., 'Cooperating in video games? Impossible! Undecidability of team multiplayer games', eds H. Ito et al., *9th International Conference on Fun with Algorithms (FUN 2018)*, Leibniz International Proceedings in Informatics (LIPIcs) no. 100 (Schloss Dagstuhl–Leibniz-Zentrum für Informatik: Dagstuhl, June 2018), article no. 14.

fully aware that the goal she or he programs the system to pursue will be followed and cannot be deviated from unless it is changed.

Depending on the goal and the capabilities provided to the system, it may follow surprising strategies (e.g. trying to control all available computational power in order to solve the shortest-route problem in order to deliver mail) and, in general, pursue unexpected instrumental goals, given enough capabilities.[13] Even considering scenarios not involving superintelligence (the appearance of which is still hypothetical), inaccurately formulated goals may result in unexpected behaviour such as a computer player exploiting a bug in a video game.[14] Notably, this behaviour occurs in the presence of adversaries with a simple behaviour; if, as mentioned above, there are agents with complex behaviours or varying, not fully cooperative, goals, then the interactions between all the agents may become very complex. In some cases, correct design of the environment (or mechanism design, using the term from economics[15]) may ensure that all agents can pursue their individual goals in a predictable fashion. However, without an explicitly designed environment, the interaction of individual, different goal functions must be carefully studied. Of course, similar problems may arise and must be studied in security-critical applications, where the consequences of unexpected behaviour are more severe.

Solving most of the above issues for general autonomous devices requires not only high-quality sensing equipment and extensive computational capacities, but also application of knowledge from two different domains of knowledge: the application area, where the autonomous system will operate, and the properties of the mathematical state estimation and control algorithms. Only when both are known is it possible to estimate how the system will operate in its environment well enough to make provably accurate predictions on the safety and correctness of its behaviour.

In the case of military applications, it is important to note that, in general, an autonomous military device must act in an adversarial environment with a complex and possibly hard-to-define goal, subject to safety and policy constraints. In such a safety-critical context as military use, the limitations of a system are potential risks, especially if they are not known beforehand and could be exploited by an adversary. A further issue specific to military use is the dual requirement of safety and predictability on the one hand and the need to defeat adversaries on the other hand, which is at least a partial conflict of goals. These issues need to be addressed at the policy level, in order to constrain a potential autonomous device's actions by a sufficiently clear and well-formulated military doctrine and a clear definition of potential use cases. This implies the need to integrate policymaking with problem domain knowledge in order to make informed and correct policy decisions and to design compliant systems.

[13] Bostrom, N., *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press: Oxford, 2014).

[14] Chrabaszcz, P., Loshchilov, I. and Hutter, F., 'Back to basics: benchmarking canonical evolution strategies for playing Atari', Computing Research Repository (CoRR), arXiv, 1802.08842, 24 Feb. 2018.

[15] Hurwicz, L. and Reiter, S., *Designing Economic Mechanisms* (Cambridge University Press: Cambridge, 2006).

IV. Conclusions

To date, recent breakthroughs in computer science have made possible the autonomization of several tasks that were previously considered to be complex, such as dependable control of a vehicle, surgery or air traffic control. However, there are limits to what computers can achieve. More complex tasks require more computational power to assess the problem domain and capture potential uncertainties and adversaries; more complex goals with additional constraints may turn out to be computationally intractable; and more complex environments need to be modelled carefully to fully capture the nuances that may radically change the decision-making process. Last but not least, the goal function of the autonomous system, which controls its decision-making process, must be modelled carefully with complete understanding of the consequences of the choice of this exact goal in the autonomous system's environment. In the military context, the conflict between safety, predictability, certification and the need to challenge the potential adversary needs to be resolved at the policy level.

None of these problems is unsolvable by itself, yet they require a deep understanding of the application domain and the mathematical foundations of the current solution methods, their capabilities and limitations, and an integration of domain and technical knowledge into doctrine design. Especially for the latter, the existing legal and political frameworks may not yield goals and constraints that can be formalized for autonomous devices; in this case, the question of whether autonomy is politically desired (and if yes, for which uses) needs to be resolved before any engineering efforts can be undertaken.

In the more specific context of nuclear risk and possible nuclear applications, the usual issues of autonomous decision-making have to be extended by the exceptionally critical nature of nuclear infrastructure. Thus, in addition to the usual engineering and computational difficulties of automation comes policy questions: To what extent is autonomous decision-making reliable? Can it be trusted to perform operations with potentially catastrophic outcomes? Is a 'human gap' a safer solution? These questions have to be posed and answered for each potential use case of autonomization.

4. Military applications of machine learning and autonomous systems

MARTIN HAGSTRÖM

This decade's renaissance in artificial intelligence (AI) has led to innovations and groundbreaking developments, as demonstrated in different applications from cars able to drive themselves to computers able to play the game go. Machine learning methods have been used to solve long-standing problems at the heart of many of these applications. It can be expected that machine learning technologies will also be applied in the development of military systems, such as autonomous weapons. Automated weapon systems have existed for more than a century and an increasing number of functions in other military systems can be expected to be automated in the future. However, for safety-critical systems such as weapons, the requirements for the development process and certification are different than for systems whose functions are less critical if an unexpected outcome occurs. Thus, it is not yet clear which machine learning methods can be used in military applications.

This essay first described the military applications of autonomous systems (section I). It then looks in particular at machine learning and its actual and potential military applications (section II).

I. Autonomy in military systems

What is an autonomous system and how does it work?

An autonomous system is a system that performs a task without human interference.[1] From a technical perspective, there is no difference between automatic and autonomous systems; the difference is mainly semantic. The term 'autonomous' is used for systems with complex automation that are able to perform complex tasks in a complex environment. An autonomous system is composed of numerous subsystems (see figure 4.1). The description is generic and can be applied to both cyber–physical systems, such as unmanned aerial vehicles (UAVs) or weapons, and decision-support systems.

To be able to act in an environment, the system needs to retrieve information about its whereabouts and its relationship to its surroundings using sensors. For airborne vehicles, the sensors measure physical conditions such as air pressure, acceleration and magnetic fields, which are used to calculate the state of the system: its velocity, altitude, direction and other variables describing the relationship of the system to its environment. These states are computed based on a mathematical model of the system and its environment. The mathematical model is a description of the interaction between the system and its universe.

[1] On the definition of autonomy and autonomous systems see chapter 2 in this volume.

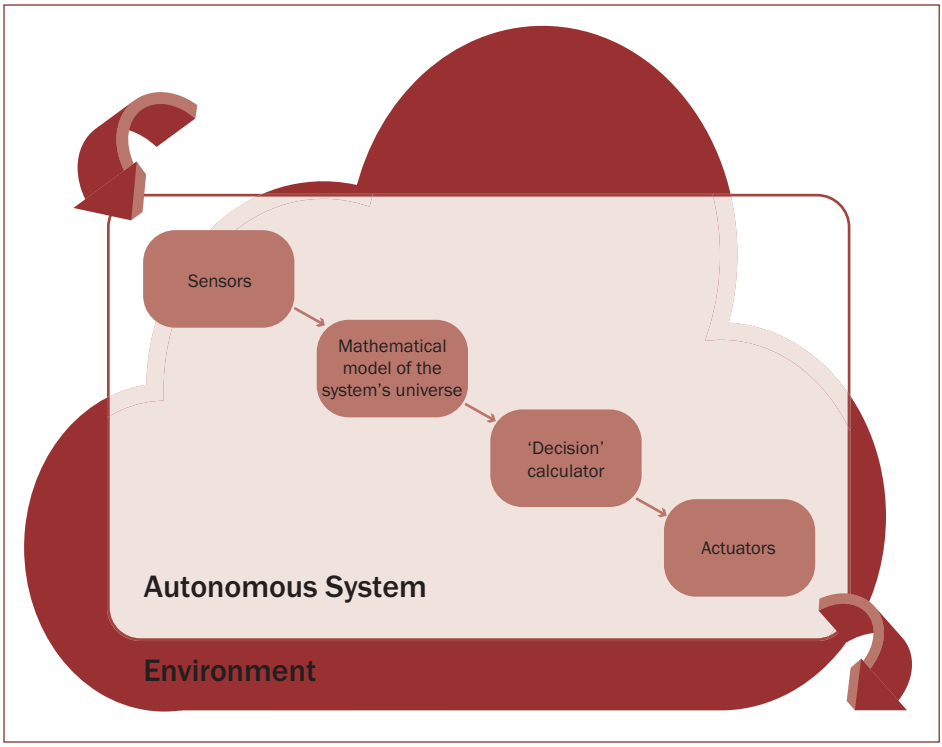

Figure 4.1. A schematic description of a generic autonomous system

Note: The system has a mathematical model of the environment, which describes the system's interactions with its surroundings.

For airborne vehicles, it describes the resulting forces on the vehicle from the surrounding air, from propulsion and from gravity. The model typically includes a description of the geographical relation to the earth and navigational states, which can be calculated using satellite navigation signals.

An 'autonomous' system always has a model of its universe, or design space, which is the mathematical model describing the system's relationship to its environment, as interpreted from the signals from the sensors. A system is intended to act only within its design space; its behaviour outside of its design space is unpredictable, as it has no description of this world.

Autonomy in weapon systems today

Automation technology is developing rapidly and is being deployed in an increasing number of applications, but it is not new.

Automation of vehicle control has been researched and developed by engineers for almost a century. The first autonomous aircraft were developed during World War I and warships were redesigned with automation for remote control

purposes in the 1930s.[2] Today, an aircraft can fly autonomously from take-off to landing. Self-driving cars have been under development for more than 20 years. As early as 1998 a self-driving car demonstrated autonomous driving along Italian highways.[3] State-of-the-art self-driving cars can travel autonomously in structured environments (but autonomous driving in a mixed environment with other road-users is still an area of research).

Depending on the definition used, autonomous weapon systems have also existed for 70 years. Guided rockets and cruise missiles were developed during World War II.[4] Today, a wide range of weapons is in use that use automation to identify, track, select and engage targets, from the Phalanx close-in air-defence system to advanced long-range cruise missiles.

Even if the automation in these systems is advanced, they are typically designed for a specific purpose within a limited design space and for a limited scope of use. The Phalanx system is self-contained in the sense that it includes a radar for target tracking and a computer to calculate target position and to aim and fire the gun.[5] The system is placed at the site it is supposed to protect and is intended to be used in automatic mode. The automatic mode is needed when incoming threats come at high speed and manual control is infeasible. To use such a system safely requires that the surroundings are clear of objects or vehicles that could be mistaken for enemy targets; there are therefore elaborate processes to follow to ensure that no unintended harm is caused before switching the system on. The human control is exercised by following procedures before using the system to ensure safe use. Human control and compliance with international humanitarian law are ensured by pre-use analysis, context-dependent use and the following of strict protocols.[6] A system with similar functionality but on a larger scale is the Aegis ballistic missile defence system.[7]

Technical and operational obstacles to advanced autonomy in military systems

Several states, with the United States in the lead, have identified increased automation as a key to future military capability.[8] It can therefore be expected

[2] Werrell, K. P., *The Evolution of the Cruise Missile* (Air University Press: Maxwell Air Base, AL, Sep. 1985), pp. 23–24.

[3] Broggi, A. et al., 'The Argo autonomous vehicle's vision and control systems', *International Journal of Intelligent Control and Systems*, vol. 3, no. 4. (1999), pp. 409–41.

[4] Werrell (note 2), p. 41.

[5] Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017), p. 38.

[6] Duke, D. S., Bahlis, J. and Morrissey, W. J., 'Evolution of maintenance training for the Phalanx Mark 15 close-in weapons system', 2008 Oxford Business & Economics Conference, 22–24 June 2008.

[7] On the safety protocols and control procedures for the Aegis see Scharre, P., *Army of None: Autonomous Weapons and the Future of War* (W. W. Norton & Co.: New York, 2018).

[8] US Department of Defense (DOD), Defense Science Board, *The Role of Autonomy in DoD Systems*, Task Force Report (DOD: Washington, DC, July 2012); US Department of Defense (DOD), Defense Science Board, *Report of the Defense Science Board Summer Study on Autonomy* (DOD: Washington, DC, June 2016); and Kalbarczyk, M., 'Autonomy in defence: systems, weapons, decision-making', *European Defence Matters*, no. 14 (Nov. 2017), p. 22.

that large efforts will be directed towards the research on and development of automation in military applications, including weapon systems. While automation has been used in weapon systems for a century, many challenges remain for the wide-ranging, generic weapon application of autonomy.

In an armed conflict a commander's decision to use force must follow the basic principles of international humanitarian law.[9] To be able to use force in a discriminate way and with necessary precaution, the behaviour and effects of weapons must be predictable. A commander must understand the effects of a weapon and be able to foresee how the weapon will behave once launched. Thus, from an operational perspective, the most notable challenge of the increasing autonomy of weapons is the need for predictability and to understand the system's behaviour.

Although autonomous systems should be independent of humans (which is what defines them as autonomous), there is always an interface between humans and systems on some level. How this interface should be designed is a distinct research field, covering the interaction between operators and the system as well as between the command structure and organization and the system.[10] The behaviour of a system with many automated functions can be difficult to comprehend and therefore the control of the system might fail, even when exercised by humans. There are examples where automated modes of systems have caused fatal accidents.[11] Designing systems that are understandable and, from the commander's view, predictable will be a challenge when introducing more complex automation (e.g. using machine learning methods).

Another challenge of automation is that any autonomous system may be vulnerable to an opponent's countermeasures. While it is mandatory that a weapon system is predictable for the user, an opponent might be able to deceive a system that has a predictable behaviour. For the same reason that it might be difficult to prove predictability of a complex system, it will be difficult to ensure that a weapon with automated behaviour has no vulnerabilities due to predictable behaviour—vulnerabilities that can be exploited by an antagonist.

II. Military applications of machine learning

What is machine learning and how does it work?

Machine learning—a collective name often used for statistical methods of identifying structures in data—has many different military applications. These methods have been used with great success to solve problems in several fields of

[9] E.g. Protocol I Additional to the 1949 Geneva Conventions, and Relating to the Protection of Victims of International Armed Conflicts, opened for signature 12 Dec. 1977, entered into force 7 Dec. 1978, Articles 57 and 58.

[10] US Department of Defense (DOD), *Unmanned Systems Integrated Roadmap FY2017–2042* (DOD: Washington, DC, 2017).

[11] Hawley, J. K., *Patriot Wars: Automation and the Patriot Air and Missile Defense System* (Center for New American Security: Washington, DC, Jan. 2017).

AI. Two problems where such methods have been famously applied to successfully solve long-standing problems are image recognition and speech recognition.

Machine learning techniques are especially well suited for data-rich applications where explicit system modelling is difficult. Every system needs a model of its universe—the system's design space—that describes the environment and the system's interactions with it. For applications where the design space is well understood and which has a mathematical description, such as describing the aerodynamic forces on an aircraft, machine learning techniques have proven to be less useful than explicit modelling (e.g. with equations relating actions and reactions between the system and the environment). In applications where there is no such concise model, but only a large set of data that implicitly describes the character of the system's universe, machine learning techniques are suitable to derive a system model. An example of such an application, where machine learning methods are commonly applied in research and development projects, is target recognition, which is, broadly speaking, an image-recognition problem.[12]

Other applications where machine learning methods are suitable and used include the following.

1. *Anomaly detection*. Machine learning methods can be used for pattern recognition. The methods can be used to identify patterns of 'normality' in data and then to detect data patterns that differ from the normal state (i.e. outliers).[13]

2. *Systems for information management in reconnaissance and surveillance applications*. Today's reconnaissance and surveillance systems collect vast amounts of information. A UAV equipped with imaging sensors can be airborne for long periods, continuously sending a stream of data through a network to an analysis centre. Analysts then assess the data and extract significant information. Data analysis can be an application where machine learning methods can prove to be useful.[14]

3. *Decision-support systems*. Decision-support systems are used in a variety of different applications such as medical diagnosis systems, manufacturing and marketing to help operators to make decisions by analysing data and proposing courses of action.

[12] Vink, J. P. and de Haan, G., 'Comparison of machine learning techniques for target detection', *Artificial Intelligence Review*, vol. 43, no. 1 (Jan. 2015), pp. 125–39.

[13] Bhatnagar, R., 'Machine learning and big data processing: a technological perspective and review', eds A. E. Hassanien et al., *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)*, Advances in Intelligent Systems and Computing no. 723 (Springer: Cham, 2018); and Liao, D. et al., 'Anomaly detection for semiconductor tools using stacked autoencoder learning', International Symposium on Semiconductor Manufacturing (ISSM), Tokyo, 10 Dec. 2018.

[14] Kuwertz, A. et al., 'Applying knowledge-based reasoning for information fusion in intelligence, surveillance, and reconnaissance', eds S. Lee, H. Ko and S. Oh, *Multisensor Fusion and Integration in the Wake of Big Data, Deep Learning and Cyber Physical System*, Lecture Notes in Electrical Engineering no. 501 (Springer: Cham, 2018), pp. 119–39; and Verma, K. et al., 'Target detection and tracking in infrared videos using frequency domain analysis and machine learning for surveillance', eds. D. Yafav et al., *5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Gorakhpur, India, 2–4 Nov. 2018 (IEEE: New York, 2018), pp. 154–59.

Machine learning in weapon systems

Machine learning methods are used for different applications in many weapon research programmes. However, there is a gap between experimental tools and fielded systems.

Before a weapon, or any military system, can be fielded it must go through extensive assessment, testing and evaluation. Since weapons in general are intended to cause harm, they can naturally cause unintended harm or accidents if they are used in an unforeseen or unintended manner or situation. Thus, it is extremely important to be able to understand and predict how a weapon system will behave. This leads the developers of weapon systems to use conservative development processes with extensive test and verification procedures.[15]

Although machine learning techniques have proven to be useful in many applications, they are not yet commonly used in weapon systems. One reason is likely to be the difficulty of the verification process of 'black box' systems, which is a characteristic of systems developed by machine learning.[16]

Technical and operational obstacles to adoption of machine learning in military systems

Challenges of both a technical and a more operational nature can be foreseen when introducing machine learning methods into weapon system development.From an operational point of view, the effects of a weapon system must be predictable to the commanding officer but with a behaviour unpredictable to the opponent. One characteristic of models created by machine learning is the statistical nature of these methods. It is still an open research problem to design models with machine learning that are transparent and whose behaviour is understandable. The models tend to become complex and the implicit design method complicates testing and verification procedures. Using machine learning while meeting military requirements of predictability and ability to understand the system's behaviour will be a challenge.

The challenges from a technological perspective generally relate to the requirement of machine learning methods for large amounts of data. Many of the applications where these methods have been used with great success are characterized by an abundance of data. The data may be collected in advance, retrieved from different sources (e.g. a database with many images of the same object in different circumstances or large sets of texts in different translations), all forming a statistical representation of the system or application to be modelled. In cases where there is no pre-recorded data representing the application, data can also be measured in relevant real-life environments (e.g. by manually driving a car) while the sensors measure and build a large data set on the environment for later use with machine learning. For some applications—including development of

[15] Defence Acquisition University (DAU), *Defense Acquisition Guidebook* (DAU: Fort Belvoir, June 2018), chapter 8.

[16] Sentient, 'Understanding the "black box" of artificial intelligence', 1 Sep. 2018.

a new weapon system—there is no pre-recorded data and it is difficult to perform real-life measurements. Data can be produced by simulation, but the simulations will depend on a simulation model and the data produced will be limited by the scope of that model. Thus, for problems that are inherently data-thin and where experiments are impractical or even infeasible, machine learning methods might not be a solution except when the application is not sensitive to this limitation.[17]

Machine learning techniques will be used where such methods are applicable, but the requirements for weapon systems are not necessarily the same as for applications where the methods have proved successful so far. It is not, therefore, obvious that they will be applied in a wide range of military systems.

III. Conclusions

Automation is not a new phenomenon, either in general or in weapons. However, military systems need to be controllable and predictable for the user and the behaviour of highly automated systems can be difficult to comprehend. The recent achievements in the field of AI have attracted a lot of attention. Many applications, including military uses, are proposed using different AI methods such as machine learning, but it is not obvious that the methods will be useful in such safety-critical applications as weapon systems. Requirements for safety and predictability must be underlined in military applications.

[17] Verma, D. et al., 'Generation and management of training data for AI-based algorithms targeted at coalition operations', eds M. A. Kolodny, D. M. Wiegmann and T. Pham, *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR IX*, 16–18 Apr. 2018, Orlando, FL, Proceedings of SPIE no. 10635 (SPIE: Bellingham, WA, 2018).

Part II. Artificial intelligence and nuclear weapons and doctrines: Past, present and future

How could recent advances in artificial intelligence (AI) have an impact in the field of nuclear weapons and doctrines? The following four essays review the extent to which AI systems could be—or already have been—used in nuclear weapon systems. Nuclear weapon systems should be understood in the broadest sense. They include not only nuclear warheads and delivery systems but also all nuclear force-related systems for nuclear command and control, early warning, and intelligence, surveillance and reconnaissance (ISR).

In the first essay of this part (chapter 5), John Borrie illustrates how the connection between AI and nuclear weapons and doctrines is not new. He shows that useful lessons could be learned from how the Soviet Union and the United States used AI and automation in their nuclear weapon systems during the cold war. The present author's second contribution to this volume (chapter 6) then discusses what could change with the current AI renaissance. It explores how recent advances in machine learning and autonomy could—theoretically—be exploited to enhance the nuclear deterrence architecture, from early-warning and command-and-control systems to nuclear weapon delivery and missile defence systems. Page Stoutland and Petr Topychkanov then assess these probabilities against the reality of the nuclear weapon modernization programmes of the USA and Russia. Stoutland (in chapter 7) reviews the role that recent advances in machine learning could play in the USA's ongoing modernization, while Topychkanov (in chapter 8) discusses how Russia sees the role of autonomy in its current and future nuclear weapon systems.

VINCENT BOULANIN

5. Cold war lessons for automation in nuclear weapon systems

JOHN BORRIE*

Dramatic advances in artificial intelligence (AI) are having wide-ranging societal impacts.[1] As part of this, concerns are being expressed about the emergence of an 'AI arms race' or an 'AI cold war that threatens us all', between the United States and China in particular.[2] Fifty years ago, the pioneering AI researcher, Marvin Minsky, described AI as 'the science of making machines do things that would require intelligence if done by men'.[3] In the near term, reality still falls a long way short of the aspiration for intelligent machines. Rather, algorithm-based machine systems are becoming vastly better at self-optimizing their performance based on various techniques, many of them related to pattern recognition and matching of data. There is potential for this to improve the ability of machine systems to perform various critical military functions with a greater level of autonomy.[4] This has led experts in forums such as the 1980 Convention on Certain Conventional Weapons (CCW Convention) to ponder the legal and moral implications of machine systems that target or attack humans without direct human supervision.[5]

The CCW discussions have been mainly concerned with robotic systems in conventional warfare.[6] However, some experts have expressed concern that, if advances in AI-related research (e.g. in machine learning) are applied to automation and increasing autonomy in nuclear early warning and command and control, these could elevate the risk that nuclear weapons will be used, for

[1] Cummings, M. L. et al., *Artificial Intelligence and International Affairs: Disruption Anticipated* (Chatham House: London, June 2018).

[2] Zwetsloot, R., Toner, H. and Ding, J., 'Beyond the AI arms race: America, China, and the dangers of zero-sum thinking', *Foreign Affairs*, 16 Nov. 2018; and Thompson, N. and Bremmer, I., 'The AI cold war that threatens us all', *Wired*, 23 Oct. 2018.

[3] Minsky, M. (ed.), *Semantic Information Processing* (MIT Press: Cambridge, MA, 1968), p. v, quoted in Geist, E. and Lohn, A. J., *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (Rand Corporation: Santa Monica, CA, 2018), p. 9. Minsky's description captures a wide-range of approaches, techniques and technologies, so its utility is limited.

[4] United Nations Institute for Disarmament Research (UNIDIR), *The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence—A Primer for CCW Delegates*, UNIDIR Resources no. 8 (UNIDIR: Geneva, 2018).

[5] Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects (CCW Convention), opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983.

[6] Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017)

* The views expressed are those of the author and do not necessarily reflect the views or opinions of the United Nations or UNIDIR's sponsors. The author thanks Dr Pavel Podvig, Kerstin Vignard and Ben Silverstein for their critical feedback on drafts of this paper. Errors are the author's own.

instance due to accidents, or have an impact on nuclear stability in various ways.[7] Yet there is much uncertainty about this. AI is a field that is rapidly evolving. Faster and more reliable, increasingly autonomous systems could, in principle, reduce the risk of nuclear weapon use in crisis situations by supporting humans to make more informed decisions.

States with nuclear weapons do not share much information about the specifics of their current—or planned—nuclear early-warning or command-and-control systems. Because of this secrecy, it is hard to judge the level or nature of the impact that more autonomous machine systems will have in practice. Nevertheless, from the limited information that is available about nuclear early-warning and command-and-control systems during the cold war, it is possible to make reasonable suggestions about how AI might affect them. Correspondingly, this essay considers what the experiences of the Soviet Union and United States in the cold war offer in considering the impact of automation and autonomy in nuclear weapon systems. It first looks in general at the use of automation in nuclear early warning and command and control by the USA and the USSR (section I), then considers the specific case of the Dead Hand automatic retaliation system (section II). Based on this historical evidence, the essay draws several conclusions about where autonomy could be taking contemporary nuclear early warning and command and control (section III). Before embarking on that, some questions of terminology are addressed.

The terms automation and autonomy are used advisedly.[8] In classic form, automated machine systems are governed by prescriptive rules that permit no deviation. Although also automated, to a greater or lesser degree autonomous systems operate without human intervention in the physical world or some kind of digital or virtual environment and select actions based on some kind of assessment of the environment's current state.[9] The actions depend on some capacity to sense and then to decide on which is most appropriate, based on algorithms.[10] During the cold war era, the technological sophistication of machine systems generally was limited compared to today, and it is difficult to recognize much autonomy in their functioning. Nevertheless, as shown below, these systems had an impact on human nuclear command-and-control decisions in ways that are telling, and which are also relevant now and in the future as AI enables more autonomous features.

One point of distinction some experts have made about increasingly autonomous systems is that, broadly, 'systems incorporating *autonomy at rest* operate virtually,

[7] E.g. Scharre, P., *Army of None: Autonomous Weapons and the Future of War* (W. W. Norton & Co.: New York, 2018), pp. 297–302.

[8] For a discussion see United Nations Institute for Disarmament Research (UNIDIR), *The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches—A Primer*, UNIDIR Resources no. 6 (UNIDIR: Geneva, 2017). On the definitions of automation and autonomy see also chapter 2 in this volume.

[9] United Nations Institute for Disarmament Research (UNIDIR), *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies*, UNIDIR Resources no. 5 (UNIDIR: Geneva, 2016). The present author was the principal author of this report.

[10] For a discussion see Boulanin and Verbruggen (note 6), pp. 5–11. The distinction between automation and autonomy is also discussed in chapters 2–4 in part I of this volume.

in software, and include planning and expert advisory systems, whereas systems incorporating *autonomy in motion* have a presence in the physical world and include robotics and autonomous vehicles'.[11] In controlling nuclear weapons it is how autonomy at rest will influence human decision-making that is of particular interest here, for reasons shown below.

Finally, although this essay is not about AI technology per se, a word is necessary about machine learning: this is an approach to increasing machine autonomy that is somewhat misleadingly termed. Algorithm-based systems do not 'learn' in a human sense. Rather, systems using these techniques can, in principle if not in practice, recursively improve their ability to successfully complete pattern recognition or matching tasks based on sets of data (which usually need to be carefully curated by humans first).[12] Such capabilities are attractive for managing and rapidly making sense of a large amount of sensory and other data, potentially including in crisis situations in which humans must make launch decisions under extreme time pressure.

I. Soviet and US use of automation in nuclear early warning and command and control

The rationale for automating nuclear command and control

Throughout the cold war, the USA and the USSR had the largest nuclear arsenals. Each developed sophisticated nuclear force-related systems for detection and early warning of nuclear attack by the other, plus command-and-control systems for its own nuclear forces.[13] The overriding imperative in the evolution of these systems was to ensure nuclear retaliatory capability in the event of an attack. This imperative still applies in the Russian and US systems almost a generation after the end of the cold war, even though the strategic and technological context has changed significantly. Today there are nine nuclear-armed states. Instead of a dyadic confrontation between the two superpowers and their respective allies as in the cold war, there are more complex possible escalation chains that may involve several nuclear-armed states in various combinations.[14] In addition, the more recent advent of technologies such as missile defences, hypersonic missiles, surface-launched anti-satellite (ASAT) weapons and offensive cyber capabilities have strategic implications that, as yet, are unclear for nuclear stability.

Nevertheless, since the USA and the USSR devoted the most attention, technology and other resources to trying to ensure that each could, if necessary, launch on

[11] US Department of Defense (DOD), Defense Science Board, *Report of the Defense Science Board Summer Study on Autonomy* (DOD: Washington, DC, June 2016), p. 5 (emphasis in original).

[12] For a basic primer see Heath, N., 'What is machine learning? Everything you need to know', ZDNet, 14 May 2018. On the definition of machine learning see chapter 2 in this volume; and on the state of the art in machine learning see chapter 4 in this volume.

[13] For general background about the evolution of the nuclear arms race see e.g. Rhodes, R., *Arsenals of Folly: The Making of the Nuclear Arms Race* (Simon & Schuster: London, 2008); Hoffman, D. E., *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy* (Anchor Books: New York, 2009); and Schlosser, E., *Command and Control* (Allen Lane: London, 2013).

[14] Krepon, M., 'Can deterrence ever be stable?', *Survival*, vol. 57, no. 3 (2015), pp. 111–32.

warning of a nuclear attack by the other (and Russia and the USA still maintain a proportion of their respective nuclear forces in this way[15]), their experiences are relevant to questions about the roles of automation and autonomy. From the 1950s, both countries built up nuclear triads consisting of crewed bombers, land-based intercontinental ballistic missiles (ICBMs) and air-launched nuclear-tipped missiles, many of which were capable of rapid launching. Tight planning, command and control were vital for this. For example, from 1960 US strategic forces operated according to a Single Integrated Operational Plan (SIOP), including pre-emptive and retaliatory options for massive nuclear attacks on targets in the China–Soviet bloc. This in itself required automated systems of various kinds, for example in-flight refuelling for bombers and the Strategic Automated Command and Control System (SACCS) to assist with logistical planning for the transmission of launch orders.[16]

The development of Soviet and US launch-on-warning postures underlined the need for automated and pre- or semi-automated systems to improve early warning in order to inform nuclear decision makers in a timely manner and to buy time for them to launch nuclear forces, if they so decided. To this end, the USA and the USSR developed sophisticated detection and early-warning capabilities based on sensors of various kinds, such as ground-based radars and dedicated early-warning satellites. They also each constructed elaborate and hardened communications, control and response systems to integrate data from various sources. Given the immense time pressure entailed in assessing whether a nuclear attack is occurring, automation was necessarily a part of some of these systems in order to ensure that attack-related information reached human decision makers.

Earlier in the cold war, the threat of enemy nuclear bombers was a main concern for both sides. To detect such attacks, each built a network of sensors such as, in the US case, the Distant Early Warning Line (DEW Line) in the 1950s. During this period, the US Air Force also trialled a computer-controlled air defence system called the Semi-Automatic Ground Environment (SAGE) to shoot down Soviet bombers over US airspace.[17] However, the USA soon shelved SAGE as the USSR deployed ICBMs. These missiles travel on ballistic trajectories largely outside the atmosphere, which makes them difficult to shoot down.[18] Moreover, the development of nuclear-powered ballistic missile submarines (SSBNs) on both sides raised the prospect of nuclear missiles coming from unexpected directions and originating nearer to targets, which would result in less time to respond. This

[15] Podvig, P., 'Risks of nuclear command and control accidents', eds J. Borrie, T. Caughley and W. Wan, *Understanding Nuclear Weapon Risks* (United Nations Institute for Disarmament Research: Geneva, 2017), pp. 53–59, p. 53.

[16] Burr, W. (ed.), *Launch on Warning: The Development of U.S. Capabilities, 1959–1979*, National Security Archive Electronic Briefing Book no. 43 (George Washington University, National Security Archive: Washington, DC, Apr. 2001).

[17] Schlosser (note 13), pp. 152–53. The SAGE system's Whirlwind computers were originally developed by the Massachusetts Institute of Technology (MIT) for the US Navy as a flight simulator.

[18] Carter, A. B., Steinbruner, S. D. and Zraket, C. A. (eds), *Managing Nuclear Operations* (Brookings Institution: Washington, DC, 1987), figure 8-4, p. 298.

necessitated the development of more sophisticated radar systems and space-based satellites as components of early-warning systems.[19]

The limits of automation

There is plenty of evidence to suggest that both the USA and the USSR recognized the limits of automation in nuclear command and control, and until the 1980s each seemed reluctant to relinquish higher-order assessment or decision-making responsibilities to automated systems outside specific situations such as missile defence.[20] They also tried to increase redundancy within their automated systems in case components were damaged, destroyed or otherwise failed. However, the increased complexity of these systems as a result of their added redundant features could also be a cause of failure.[21]

In fact, early-warning systems on both sides suffered numerous faults and false alarms or were simply too constrained in their automatic capabilities to be fully reliable.[22] A major role of humans 'in the loop' in nuclear command-and-control systems was to respond to these problems as they arose. Three examples illustrate this.[23]

1. In November 1979 a war exercise tape loaded mistakenly onto a computer at the North American Aerospace Defense Command (NORAD) fed data on an incoming nuclear attack into the early-warning system. Only NORAD's ability to independently check its radar system (a practical example of redundancy) revealed to operators that this was a false alarm. This reflected a US approach of 'dual phenomenology'—having multiple, independent forms of tactical sensor for comparison. However, this redundancy was expensive and was not always present in the Soviet early-warning system.[24]

2. At 02.26 on 3 June 1980 US military commanders at NORAD telephoned the US president's national security advisor, Zbigniew Brzezinski, to tell him that 220 ICBMs were inbound from the USSR. NORAD commanders called back a few minutes later to increase the estimate to 2200 nuclear missiles. Brzezinski prepared himself to tell President Jimmy Carter, who would only have a few

[19] eds Carter et al. (note 18), table 8-2, p. 311, and table 8-3, p. 313.

[20] E.g. the USA had the Perimeter Acquisition Radar Attack Characterization System (PARCS) to help characterize attacks and probable nuclear warhead impact points. Bethmann, R. C. and Malloy, K. A., 'Command and control: an introduction', Master's thesis, Naval Postgraduate School, Mar. 1989, p. 83.

[21] Perrow, C., *Normal Accidents: Living with High-Risk Technologies*, 2nd edn (Princeton University Press: Princeton, NJ, 1999).

[22] On the faults and false alarms see Sagan, S. D., *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons* (Princeton University Press: Princeton, NJ, 1993); and Lewis P. et al., *Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy* (Chatham House: London, Apr. 2014), p. 7. On reliability in the USA see Carter et al. (note 18). On reliability in the USSR see Podvig, P (ed.), *Russian Strategic Nuclear Forces* (MIT Press: Cambridge, MA, 2001).

[23] These examples are from Lewis et al. (note 22). See also Borrie, J., 'A limit to safety: risk, "normal accidents", and nuclear weapons', International Law and Policy Institute–United Nations Institute for Disarmament Research Vienna Conference Series no. 3, Dec. 2014.

[24] Podvig, P., 'No gaps in early-warning coverage as three radars to begin combat duty in 2017', 23 Dec. 2016, Russian Strategic Nuclear Forces.

minutes to decide whether to launch nuclear retaliation. A short time later Brzezinski received a third call: the attack was a false alarm. Technicians at the NORAD command centre eventually found that the incident had been cause by the failure of a computer chip costing less than US$1.

3. In September 1983 the Soviet early-warning system reported five ICBMs inbound from the USA. The watch officer on duty, Lieutenant Colonel Stanislav Petrov, had to make sense of the warning. Petrov was sceptical that it was really a nuclear attack: why would the USA attack with only five ICBMs when it must know that the Soviet nuclear retaliation would be massive? Indeed, it turned out that what the system took for missile plumes was a reflection off clouds. Although human judgement and capacity for contextual thinking is by no means infallible, in this case it was key to correctly assessing the threat.

II. Automation and the Dead Hand

In principle, the USA and the USSR had roughly similar nuclear early-warning and command-and-control systems, at least in terms of their objectives. In practice, these systems were not precisely symmetrical for both organizational and technical reasons. Soviet technology tended to lag, especially later in the cold war as the USSR struggled to put in orbit a satellite-based launch-detection constellation that was as capable or durable as the US system. Nor was the Soviet ground-based detection network as extensive in its coverage.[25]

In the first half of the 1980s, following the 1979 Soviet invasion of Afghanistan, the USA built up its military forces. US President Ronald Reagan used bellicose language about the USSR constituting an 'evil empire' and proposed a missile defence system to protect the USA from Soviet nuclear attack. The USSR feared that, if such missile defences came to pass, a US first nuclear strike could decapitate Soviet nuclear command and control, and thus destroy most of its nuclear forces before they could launch, with US missile defences mopping up the remnants that did.[26] Anxiety among Soviet policymakers about assuring their nuclear retaliatory capability led them towards an alternative option that depended on a greater level of automation—the Mertvaya Ruka (Dead Hand) system, which the USSR brought online in 1985.[27]

There are differing accounts as to the functioning of Dead Hand.[28] It is sometimes confused with the Perimetr (Perimeter) system, an automatic system of signal rockets used to beam radio messages to launch nuclear missiles if other means of communication were knocked out. The Dead Hand system could have used Perimetr, but the two are not exactly the same.[29]

[25] Podvig, P., 'History and the current status of the Russian early-warning system', *Science and Global Security*, vol. 10, no. 1 (2002), pp. 21–60.

[26] Thompson, N., 'Inside the apocalyptic Soviet doomsday machine', *Wired*, 21 Sep. 2009.

[27] Thompson (note 26).

[28] Podvig, P., UNIDIR, Conversation with author, 27 Nov. 2018.

[29] Podvig (note 22), pp. 65–66. A different interpretation is presented in chapter 8 in this volume.

Dead Hand was a feature of the Soviet command-and-control system designed to ensure nuclear retaliation in case of an attack, and by some accounts it is still operational in Russia today.[30] The system allowed the Soviet command authority to issue a preliminary command to its nuclear forces to enable them to accept a launch order. Normally, the command authority would generate the actual launch order, which the command-and-control system would carry out once a certain set of conditions were met. These conditions could include seismic, radiation and air pressure data indicating nuclear explosions from a network of sensors.

The Dead Hand system was also apparently able to operate in a semi-automated mode that did not require the order from the command authority to launch an attack. Before issuing that kind of launch command, the system would also have to check that all conditions were met. First the preliminary command was generated. Then the system would determine if a nuclear weapon had struck the USSR. If it seemed that one had, the system would presumably check to see if communication links to the Soviet command authority remained. If those links were down, then the system would infer that a nuclear attack had occurred. It would immediately transfer launch authority to whoever was manning the system at that moment deep inside a protected bunker—bypassing many layers of normal command authority. At that point, the ability to launch nuclear attack would fall to whichever small group of officers was on duty. They would now have the authority to launch the Perimetr system to communicate launch orders to silos around the USSR, as well as to submarines and bombers.[31]

The USSR apparently believed the system would add to stability because it meant that its leaders would not have to launch prematurely under pressure in a crisis situation. Since it guaranteed nuclear retaliation, they could afford to switch on the system and wait. In retrospect, the Dead Hand system resembles the Soviet doomsday machine in the classic 1964 film *Dr. Strangelove*.[32] As in the film, the USSR did not tell the USA about the system—even though it might have had a deterrent effect.[33] Meanwhile, although the USA also had a Perimetr-like Emergency Rocket Communications System, it was never combined into a system analogous to Dead Hand out of fear of accidents that might lead to nuclear catastrophe.[34]

The Dead Hand system was semi-automated. While a human finger was ultimately on the nuclear button or key, it was the culmination of a chain of developments in Soviet command and control that was by then 'ultrafast and

[30] Thompson, N., 'The Soviets built a doomsday machine. It's still working', *Wired*, 22 Sep. 2009. See also chapter 8 in this volume.

[31] Thompson (note 26).

[32] The film *Dr. Strangelove* includes this exchange between Dr Strangelove and the Russian ambassador: 'Dr Strangelove: Of course, the whole point of a Doomsday Machine is lost, if you *keep* it a *secret*! Why didn't you tell the world, EH? Ambassador de Sadesky: It was to be announced at the Party Congress on Monday. As you know, the Premier loves surprises.' Quoted in Podvig, P., 'Dr. Strangelove meets reality', Russian Strategic Nuclear Forces, 14 Apr. 2006.

[33] Hoffman (note 13), p. 154.

[34] Thompson (note 26).

largely automated'.[35] One way of looking at the human decision makers in the system—the duty officers—is that

> [They] are just another cog in an automatic, regimented system. If the duty officers are drilled over and over again to follow the checklist, and if the highest authorities had given the permission from the top, and if all three conditions on the checklist are met, wouldn't they naturally do as they had been trained to do?[36]

Nor was this necessarily peculiar to the USSR. The 1980 NORAD 'computer chip' incident described above shows that in the USA military duty officers would go through all the motions, even when it arguably should have been clear that something was wrong with the system. Bruce Blair, a former US Air Force officer, has argued that this risk is still present in the Russian and US nuclear command-and-control systems.[37]

III. Where could autonomy be taking nuclear early warning and command and control?

Increasing reliance on automation for early warning and target detection

Dead Hand was not an autonomous weapon, let alone anything 'intelligent'. Governed by simple if–then conditions, it was more like an automated telephone exchange. Automation has come a long way since the 1980s, not least because far more computational processing power is now available. While these and other advances may enable higher levels of machine autonomy in some situations, machines still struggle with contextual thinking. This remains a thorny challenge for AI research, let alone for practical application.[38] Nevertheless, the reality is that automation has a large and continuously accruing impact on sensing, surveillance, analysis and many other functions related to nuclear command and control, even if recent public attention has focused on the antiquated nature of some features of the Soviet and US nuclear force command systems. (For example, in 2016 it was revealed that SACCS, originally fielded in 1963, was still running on floppy disks, a 1970s technology.[39])

It would therefore be surprising if machine learning techniques are not being applied already to specific nuclear problems associated, for instance, with detection and early warning and with target identification. The US Department of Defense already applies such techniques to triage and process data sets through undertakings such as Project Maven, which brings AI—specifically, deep neural

[35] Hoffman (note 13), p. 153.

[36] Hoffman (note 13), p. 153.

[37] Blair, B. G., *The Logic of Accidental Nuclear War* (Brookings Institution: Washington, DC, 1993), p. 181.

[38] For an accessible overview see Thompson, C., 'How to teach artificial intelligence some common sense', *Wired*, 13 Nov. 2018. On the technical limitations of AI technology see chapters 2–4 in part I of this volume.

[39] US Government Accountability Office (GAO), *Information Technology: Federal Agencies Need to Address Aging Legacy Systems*, GAO-16-468 (GAO: Washington, DC, May 2016), p. 60.

networks—to the fight against the Islamic State group.[40] According to one recent study, as such techniques advance in the nuclear context, prospective (if not actual) capabilities in machine learning and other AI-related techniques could affect assured nuclear retaliatory capability and thus current strategic balances.[41] In the view of that study's authors, 'rapid technical progress in AI and its many potential intersections with nuclear strategy' mean that this challenge is acute, listing progress on capabilities such as analysis of intelligence, surveillance and reconnaissance (ISR) data, controlling autonomous sensor platforms, and automated target recognition.[42]

Changing impacts on human decision-making processes

How these systems will, in effect, contribute to shaping human decision makers' perceptions is not a trivial problem. Lora Saalman of the EastWest Institute, for instance, has observed that China is focusing on AI as a key technology and has a distinctly different viewpoint to the West on the roles, benefits and risks of machine learning for its nuclear forces.[43] In addition, there are the problems of automation 'surprises' and 'bias'. Nasty automation surprises could occur simply because operators of a system cannot diagnose and respond as quickly as is necessary to prevent an unsafe deviation from the intended activity (e.g. warning of inbound ICBMs).[44] Automation bias (i.e. complacency or over-reliance on the automated or autonomous system) has been a cause of accidents across a range of fields, from aviation to clinical decision-support systems used in medicine.[45] The problem here is not limited to the machine part of the system; it is the way that human operators interpret and rely on them. In a nuclear crisis situation, will there be time to check?

In the absence of declassified information about current nuclear early-warning and command-and-control systems, it is difficult to assess the pros and cons of AI-enabling aspects of these systems. One way to think about the extraordinarily complex command, control, communications and intelligence (C3I) systems for nuclear weapons is that they have a dumb-bell shape: two circles, connected by a bar. The left-hand circle, or weight, is the detection and early-warning system. The right-hand circle is the post-decision response. As described above, activities and processes falling within each of these circles are often highly automated. In terms of response, for instance, missiles cannot be recalled once launched from

[40] Pellerin, C., 'Project Maven to deploy computer algorithms to war zone by year's end', US Department of Defense, 21 July 2017. See also Allen, G. C., 'Project Maven brings AI to the fight against ISIS', *Bulletin of the Atomic Scientists*, 21 Dec. 2017; and Lynch, J., 'Why the Air Force is investing $100M in AI', Fifth Domain, 6 Dec. 2018.

[41] Geist and Lohn (note 3).

[42] Geist and Lohn (note 3), pp. 8–9.

[43] Saalman, L., 'Fear of false negatives: AI and China's nuclear posture', *Bulletin of the Atomic Scientists*, 24 Apr. 2018.

[44] United Nations Institute for Disarmament Research (note 9), p. 12.

[45] Cummings, M., 'Automation bias in intelligent time critical decision support systems', American Institute of Aeronautics and Astronautics 1st Intelligent Systems Technical Conference, Chicago, IL, 20–22 Sep. 2004, pp. 1, 5.

silos and submarine hatches: it is difficult enough to recall crewed bombers. The bar connecting the two weights represents assessment and decision. Until now, so far as is known, this has been a domain of human judgement and decision. This is true even of the Dead Hand if it is considered as a sort of pre-delegation system.

An important question to consider is what the impact of increasingly sophisticated algorithm-based systems is on this bar in the C3I dumb-bell. Do the weights get closer or further away from one another? What happens if it gets to the point that the two weights touch; that is, algorithm-based systems extend all the way into the assessment and decision-making phase? This is an important question because numerous experts have warned that it is vital that detection and early-warning systems of nuclear attack are independent of other parts of the nuclear command-and-control chain. Among them, Charles Perrow, Scott Sagan and Paul Bracken have each cautioned that hidden interactions and 'system accidents' can result in unexpected and potentially catastrophic outcomes in nuclear command-and-control systems.[46] The weights should not touch.

As machine learning and other AI techniques that may seem like a black box to operators permeate human assessment and decision-making, will the C3I dumb-bell bar lengthen or shorten? Will human decision makers be able to retain the contextual awareness to allow them to make correct decisions? And then to what extent is retaining humans 'in the loop' still a safeguard? Bruce Blair has argued of the existing US ICBM launch system that airmen sitting in a bunker following preset instructions to launch if certain conditions are met are really part of an automated system.[47] If an AI-enabled system does not permit an operator meaningful re-assessment of a situation, or discourages it due to automation bias, it could be as bad as the case where human beings are just cogs in a largely automatic, regimented system like Dead Hand.

IV. Conclusions

Three main conclusions can be drawn from this brief selection of Soviet and US cold war experiences.

First, from early in the nuclear age both the USA and the USSR grappled with the questions of which assessment and decision-making roles are appropriate for delegation to machines and what is an appropriate level of delegation. The promise of new automated technologies was either countered by the adversary (e.g. SACCS versus ICBMs) or frequently could not achieve the level of performance needed for decision makers to be fully confident in its reliability (as shown by the three examples in section I). In general, nuclear decision makers in both states seemed to be deeply aware that, when dealing with something as tightly-coupled, complex and potentially hazardous as nuclear command and control, machine-based systems face real limits that require meaningful human control and supervision.

[46] Perrow (note 21), especially chapter 8; Sagan (note 22); and Bracken, P., 'Instabilities in the control of nuclear forces', ed. M. Hellman, *Breakthrough: Emerging New Thinking—Soviet and Western Scholars Issue a Challenge to Build a World Beyond War* (Walker and Co: New York, 1988), pp. 21–30.

[47] Blair (note 37), p. 181.

Yet this did not prevent the USSR from developing the Dead Hand system, driven by a desire to give confidence to its nuclear decision makers that nuclear retaliation remained assured in a crisis.

Second, whether increasingly automated or autonomous systems elevate or reduce nuclear risk will depend on a number of factors. These factors include how system designers implement autonomous functions—itself an inherently value-laden process and not necessarily a fully rational one[48]—and how both operators and potential adversaries understand their capabilities and limitations. In the case of the Dead Hand system, the adversary was not even aware that it existed. On that specific basis, at least, the USSR could not expect the USA to display the kind of caution that the USSR presumably desired in terms of averting risky strategic behaviour.

The international security situation has lately deteriorated and strategic rivalry has intensified among several of the nuclear-armed states. All of these states are modernizing their nuclear systems, and some are considering additional roles for their nuclear forces or announcing new capabilities. In such circumstances, misperception or misunderstanding could bring about false alarms or nuclear crises. It is conceivable that increasingly autonomous or AI-enabled decision-support systems that are intended to provide a clearer real-time picture to decision makers might have the opposite effect.

In the light of this possibility, a third conclusion is that it would be prudent for each nuclear-armed state to ensure that it understands the role of automation and autonomy in the nuclear early-warning and command-and-control systems of the others, as well as the constraints of its own systems. The USA and the USSR spent a great deal of time and effort studying each other's strategic systems and behaviour during the cold war and their military representatives met frequently, even if not always productively. As AI is integrated into military systems in general in the coming years, each nuclear power may do it in different ways, as, for instance, Lora Saalman indicated by contrasting how China and the West understand concepts related to AI and nuclear deterrence.[49] This underlines the importance of regular military-to-military contacts on the matter, perhaps as adjuncts to contacts on maintaining stability and planning for crisis management.

A striking point about the Soviet officer Stanislav Petrov in 1983 was his healthy scepticism about the reliability of the technology that he was dealing with. Can this be counted on in the current era, when remarkably high levels of reliability of devices such as smartphones, rapidly improving virtual assistants such as Siri and Alexa, and even autonomous features in vehicles—that would have seemed wondrous just a generation ago—are taken for granted now? Some experts have suggested not.[50]

[48] United Nations Institute for Disarmament Research (UNIDIR), *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies—A Primer* UNIDIR Resources no. 9 (UNIDIR: Geneva, 2018).

[49] Saalman (note 43).

[50] E.g. Hayes, P., 'Nuclear command-and-control in the millennials era', Nautilus Institute for Security and Sustainability, 17 Feb. 2015.

For the near future, it is hard to see a situation in which humans explicitly delegate decisions to launch nuclear forces to machines—although the Soviet experience in the 1980s indicates that the possibility of movement in that direction should not be discounted. For all of the extended capabilities that AI-enabled systems may offer nuclear early warning and command and control, nuclear policymakers and operators need to keep at the forefront of their minds the question of what this helping hand could take away in the process if it is not implemented well and under meaningful human control.[51]

[51] For a discussion of meaningful human control see United Nations Institute for Disarmament Research (UNIDIR), *The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward*, UNIDIR Resources no. 2 (UNIDIR: Geneva, 2014).

6. The future of machine learning and autonomy in nuclear weapon systems

VINCENT BOULANIN

The field of nuclear weapons is renowned for its conservativeness. For safety and security reasons, it has been slow to integrate some of the major developments in information and communications technology, as they could introduce new vulnerabilities or reduce reliability. This is particularly the case for nuclear command and control, which continues to rely on obsolete cold war technology. The US military, for instance, still uses 8-inch floppy disks to coordinate nuclear force operations.[1]

Russia, the United States and a number of other nuclear-armed states have declared their intention to modernize their nuclear command-and-control systems by retiring some of these legacy systems and adopting state-of-the-art digital technologies.[2] In most cases there is no end date associated with the modernization plans, and it is hard to predict when and how the transition from cold war-era technology will take place. It is not difficult to imagine, however, that nuclear-armed states will try to make use of the current renaissance in artificial intelligence (AI). The question then is: what could the impact be? However, since AI and automation have been part of the nuclear deterrence architecture for decades, the recent advances in AI may not have a transformative impact at all.

This essay reviews how recent advances in machine learning and autonomy might be used in nuclear weapon systems and discusses the extent to which these potential applications might differ from how AI and automation have historically been used. It looks at four key areas of the nuclear deterrence architecture: early warning and intelligence, surveillance and reconnaissance (ISR) in section I; command and control in section II; nuclear weapon delivery in section III; and non-nuclear operations in section IV.

I. Early warning and intelligence, surveillance and reconnaissance

Machine learning and autonomy hold major promise for early warning and ISR. The potential of machine learning is this area derives from three abilities.

1. *Making early-warning and ISR systems more capable*. Machine learning can be used to give any type of ISR system more perceptual intelligence. One foreseeable development would be a mobile ISR platform (e.g. a surveillance

[1] US Government Accountability Office (GAO), *Information Technology: Federal Agencies Need to Address Aging Legacy Systems*, GAO-16-468 (GAO: Washington, DC, May 2016), p. 60.

[2] On these modernization plans see chapters 7 and 8 in this volume. On the current status of the nuclear modernization programmes of nuclear-armed states see Kile, S. N. et al., 'World nuclear forces', *SIPRI Yearbook 2018: Armaments, Disarmament and International Security* (Oxford University Press: Oxford, 2018), pp. 235–87.

drone) that could process data on-board and identify by itself not only signals or objects but also situations of interest such as unusual movement of troops. A number of ongoing experimental research projects aim to develop these types of capability for conventional weapons. A notable example is the Automated Image Understanding project of the US Office of Naval Research, which is intended to develop techniques to infer intentions and threats from surveillance imagery.[3] These capabilities could be repurposed for nuclear-related ISR.

2. *Searching and making sense of large sets of intelligence data*. Machine learning can be used to find correlations in large and potentially heterogeneous sets of intelligence data. An early illustration is the US military's Project Maven, also known as the Algorithmic Warfare Cross-Function Team, which aims to use machine learning to automatically analyse video surveillance footage gathered during counterinsurgency operations in Iraq, Afghanistan and elsewhere.[4] The next step for the US military is to look for correlations in different types of data set.[5] This type of capability is currently mainly pursued for counterterrorism purposes, but it is not hard to imagine that it could also be useful for nuclear-related early-warning and ISR missions, as it would permit the military commander to have better situational awareness.

3. *Making predictions*. Data-processing capability can be used to help the military command to predict developments related to nuclear weapons, including the possible production, commissioning, deployment and use of nuclear forces by adversaries.[6] The cross-analysis of intelligence data using machine learning algorithms could help the military to identify more quickly and reliably if a nuclear attack is or could be under way.

In sum, machine learning could give the human military command better situational awareness and potentially more time to make decisions.

The primary value of autonomy and autonomous systems is that they could improve the remote-sensing capabilities of nuclear-armed states—be it for early-warning or nuclear ISR missions. The main advantages of autonomous systems compared to remotely controlled and manned systems are that they can achieve greater reach, persistence and mass: they can be safely deployed in such operational theatres as deep water or areas protected by anti-access/area-denial (A2/AD) systems; they can conduct extended mission over days or, in the case of

[3] US Office of Naval Research, 'Computational methods for decision making—automated image understanding', [n.d.].

[4] Weisgerber, M., 'General: Project Maven is the just the beginning of the military's use of AI', Defense One, 28 June 2018. The project is reportedly due to end in 2019. Wakabayashi, D. and Scott, S., 'Google will not renew Pentagon contract that upset employees', *New York Times*, 1 June 2018. On Project Maven see also chapters 2, 5, 6, 10 and 11 in this volume.

[5] US Department of Defense (DOD), Defense Science Board, *Report of the Defense Science Board Summer Study on Autonomy* (DOD: Washington, DC, June 2016).

[6] US Department of Defense (note 5).

underwater systems, even months; and they can potentially be deployed in great number as they can be relatively inexpensive.[7]

These attributes are particularly attractive in the conduct of nuclear-related ISR operations, particularly submarine reconnaissance. Many types of autonomous platform could be used for this type of mission including autonomous vessels (also known as autonomous surface vehicles, ASVs), autonomous underwater vehicles (AUVs) and autonomous aerial vehicles (AAVs). The USA has already developed a prototype ASV, *Sea Hunter*.[8] The USA, Russia, China, Japan and a few other states are also developing autonomous underwater systems. Systems such as the US Littoral Battlespace Sensing-Gilder programme may be manufactured at a relatively low cost and, thus, deployed on a massive scale.[9] In the case of AAVs, existing large unmanned aerial vehicles (UAVs) such as the RQ-4 Global Hawk could be used for this type of mission.[10]

The extent to which the deployment of such systems will change submarine warfare has been debated. On the one hand, some experts believe that, if deployed in great numbers, these platforms would make at-sea deterrence obsolete.[11] On the other hand, some believe that the potential of these systems is overstated, given that (*a*) few of the sensors carried by these systems would be able to detect deeply submerged submarines, (*b*) the range of these sensors is limited and (*c*) nuclear-powered ballistic missile submarines (SSBNs) carrying nuclear weapons operate over vast areas, so the chance of detection is negligible, even if many autonomous reconnaissance systems were deployed.[12] Major advances in power, communications and sensor technologies would be needed before these systems can have a revolutionary impact on submarine reconnaissance. Nonetheless, they could play an important support role in anti-submarine warfare. If deployed at choke points or the enemy's exit routes, these systems could, for instance, serve as a virtual barrier that would deter or deny an opponent's submarines the ability to operate in specific areas.[13]

II. Command and control

In the near term, recent progress in machine learning and autonomy is unlikely to have a major transformative impact on nuclear command-and-control systems. There are two reasons for this. First, command-and-control systems already

[7] On the benefits of autonomy see chapter 2 in this volume; and Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017). On A2/AD systems see Simon, L., 'Demystifying the A2/AD buzz', War on the Rocks, 4 Jan. 2017.

[8] US Defence Advanced Research Projects Agency (DARPA), 'ACTUV "Sea Hunter" prototype transitions to Office of Naval Research for further development', 30 Jan. 2018.

[9] Teledyne Brown Engineering, 'LBS-G: Littoral Battlespace Sensing–Gliders', Apr. 2018.

[10] On the potential use of UAVs for nuclear-related mission see chapter 12 in this volume.

[11] Hambling, D., 'The inescapable net: unmanned systems in anti-submarine warfare', British–American Security Information Council (BASIC) Parliamentary Briefings on Trident Renewal no. 1, Mar. 2016. See also chapters 9, 10 and 14 in this volume.

[12] Gates, J., 'Is the SSBN deterrent vulnerable to autonomous drones?', *RUSI Journal*, vol. 161, no. 6 (2016), pp. 28–35.

[13] Gates (note 12).

rely on a great degree of automation. Second, the types of algorithm underlying machine learning-driven applications and complex autonomous systems remain too unpredictable due to the problems of transparency and explainability.[14] Nuclear command-and-control systems are too safety-critical to be left to algorithms that engineers and operators cannot fully understand. Moreover, relatively traditional rule-based algorithms would be sufficient to further automate command and control—however, there seems to be a general agreement among nuclear-armed states that this should not be done, even if technological developments would permit it.[15]

Even if they are not transformative, advances in machine learning and autonomous systems could bring some qualitative improvements in the nuclear command-and-control architecture. They could be used to enhance protection against cyberattacks and jamming attacks. Machine learning could also help planners to more efficiently manage their forces, including their human resources. Similarly, autonomous systems could be used to enhance the resilience of the communications architecture. Long endurance UAVs could, for instance, be used to replace signal rockets in forming an alternative airborne communications network in situations where satellite communication is impossible.

III. Nuclear weapon delivery

Advances in machine learning and autonomy are likely to have an impact on nuclear weapon delivery in different ways.

In the case of machine learning, the impact is likely to result primarily in a qualitative improvement in the delivery systems. Machine learning could be used to make nuclear delivery systems capable of navigating to their target more autonomously and precisely (with less reliance on humans setting navigation and guidance parameters). A number of countries are currently exploring the use of machine learning to develop control systems for hypersonic vehicles, which, because of their high velocity, cannot be operated manually.[16] It could also make them more resilient to countermeasures such as jamming or spoofing.

In the case of autonomy, systems such as UAVs, and in particular unmanned combat aerial vehicles (UCAVs), and unmanned underwater vehicles (UUVs) could have a more transformative impact than machine learning since they provide an alternative to manned bombers and manned submarines as well as intercontinental ballistic missiles (ICBMs). Their comparative advantages include their extended endurance and their recoverability.[17]

[14] On problems of transparency, explainability and predictability of machine learning systems see chapters 2 and 4 in this volume.

[15] On the use of automation in command and control during the cold war see chapters 5, 7 and 8 in this volume.

[16] Saalman, L., 'China's integration of neural networks into hypersonic glide vehicles', ed. N. D. Wright, *AI, China, Russia, and the Global Order: Technological, Political, Global, and Creative Perspectives*, White Paper (US Department of Defense and Joint Chiefs of Staff: Washington, DC, Dec. 2018), pp. 153–60.

[17] On UCAVs in particular see chapter 12 in this volume.

Unmanned vehicles—whether remotely controlled or autonomous—can conduct much longer missions than their manned counterparts. This is particularly notable for unmanned aircraft, which can stay in flight for several days, particularly if in-flight refuelling or the use of solar power is possible. The endurance record for an unmanned aircraft of 26 days was set by a solar-powered UAV from Airbus in 2018.[18] Increased endurance also means greater reach: an unmanned platform can cover a much larger area and, in the case of an underwater system, reach greater depths than a manned vehicle. The extended endurance of unmanned platforms potentially increases their ability to survive countermeasures. A UUV, for instance, would rarely, if ever, have to return to port, which would make it harder to find and track. Combined, these benefits could, arguably, decrease policymakers' fear of a nuclear decapitation.[19]

The recoverability of UAVs and UUVs also sets them apart from missiles and torpedoes and offers policymakers new tools for managing escalation in a crisis or conflict. The decision to launch an unmanned system on patrol is not equivalent to the decision to launch a one-way device such as a nuclear ICBM or torpedo (although some such systems may be aborted after launch). Recoverability gives decision makers greater flexibility in that they would have more time to make a decision and, potentially, to recall the system.

The added value of autonomous systems lies, in other words, less in the degree of automation or autonomy but in the physical properties of robotics platforms. ICBMs and SLBMs, once launched, already operate de facto autonomously since they rely on automation to set their flight trajectory and navigate to their target. While autonomy enhances the strategic value of robotics platforms, it is not an essential requirement (with the notable exception of underwater systems, which cannot be operated remotely).

At least two nuclear-armed states are considering the possibility of using UAVs or UUVs for nuclear delivery. In 2015 a Russian television report revealed that Russia was developing a large nuclear-armed UUV, Poseidon (previously known as Status-6).[20] The system, which has been described as both a long-range torpedo and an unmanned submarine, reportedly has a range of 10 000 kilometres and a speed of 56 knots and can descend to a depth of 1000 metres.[21] It will operate autonomously but, as explained above, that is primarily a requirement of its operating environment. The USA is also building a nuclear-capable bomber, the B-21 Raider, which would reportedly be 'optionally-manned'.[22] The USA

[18] Airbus, 'Airbus Zephyr Solar High Altitude Pseudo-Satellite flies for longer than any other aircraft during its successful maiden flight', 8 Aug. 2018.

[19] Scharre, P., Horowitz, M. C. and Velez-Green, A., 'A stable nuclear future? The impact of automation, autonomy, and artificial intelligence', Working paper, University of Pennsylvania, 2017.

[20] Oliphant, R., 'Secret Russian radioactive doomsday torpedo leaked on television', *Daily Telegraph*, 15 Nov. 2015.

[21] Insinna, V., 'Russia's nuclear underwater drone is real and in the Nuclear Posture Review', *Defense News*, 12 Jan. 2018. On the system and the rationale behind its development see chapter 8 in this volume. It is also discussed in chapters 9, 11 and 14 in this volume.

[22] Majumdar, D., 'USAF leader confirms manned decision for new bomber', *Flight International*, 23 Apr. 2013. See also Gates, R., US Secretary of Defense, 'Statement on department budget and efficiencies', US Department of Defense, 6 Jan. 2011.

has not specified whether it would be prepared to operate the bomber remotely while carrying nuclear weapons, but a 2013 US Air Force report suggests that is unlikely: 'Certain missions [for unmanned aircraft], such as nuclear strike, may not be technical feasible unless safeguards are developed and even then may not be considered'.[23] It is thus hard to imagine that the USA is currently considering the use of autonomously piloted UAVs for nuclear weapon delivery. That being said, the technology exists. Existing prototypes of UCAVs (including the Northrop Grumman X-47B, the Dassault nEUROn and the BAE Systems Taranis) could—theoretically—be used for nuclear strikes.

IV. Non-nuclear operations

Nuclear-armed states, and also non-nuclear-armed states, could use machine learning and autonomy in non-nuclear applications with a strategic effect.

Missile, air and space defences

Machine learning methods could significantly improve the targeting capability of conventional defensive systems. Missile and air defence systems have relied on automation for decades. The first automatic air defence system, the Mark 56 gun fire-control system, was invented during World War II.[24] Since the 1970s, air defence systems have been using an AI technology known as automatic target recognition (ATR) that can detect, track, prioritize and select incoming air threats more rapidly and more accurately than a human possibly could. However, the progress of the target-identification capabilities of these systems has been slow, particularly due to the difficulties associated with the development of target libraries (i.e. the database of target signatures that an ATR system uses to recognize its target).

With traditional AI programming methods, the designers of an ATR system have to upload a large and representative sample of data about the target in all conceivable variations of its operating environment (i.e. background and weather conditions). This is a challenging task for many target types and operational situations.[25] Advances in machine learning, particularly deep learning and generative adversarial networks (GANs), could significantly simplify that process.[26] With deep-learning methods, engineers could make ATR systems capable of learning independently not only the differences between types of target but also the differences between military and civilian objects (e.g. a commercial

[23] US Air Force, *RPA Vector: Vision and Enabling Concepts, 2013–2038* (Headquarters US Air Force: Washington, DC, 17 Feb. 2014), p. 54.

[24] Mindell, D. A., 'Automation's finest hour: radar and system integration in World War II', eds A. C. Hughes and T. P. Hugues, *Systems, Experts and Computers: The Systems Approach in Management and Engineering, World War II and After* (MIT Press: Cambridge, MA, 2000), pp. 27–56, pp. 40–44.

[25] Ratches, J. A., 'Review of current aided/automatic target acquisition technology for military target acquisition tasks', *Optical Engineering*, vol. 50, no. 5 (July 2011), article no. 072001.

[26] On deep learning and GANs see chapter 2 in this volume.

aeroplane and a strategic bomber).[27] With GANs, engineers could generate realistic synthetic data on which an ATR system can be trained and tested in simulation. An ATR system trained with these machine learning techniques would perform comparatively much better than an ATR system trained with traditional methods.

Equally, autonomous systems offer new defensive tools against incoming threats. Autonomous unmanned vehicles can be deployed as decoys or flying mines to complement traditional air defences.[28] Advances in autonomy for swarming and for multi-vehicle control could also enable autonomous unmanned systems to operate in a coordinate way and conduct advanced A2/AD manoeuvres.[29] Such systems would increase deterrence against both conventional and nuclear attack as they would increase the risks for an attack by manned platforms (e.g. combat aircraft and manned bombers) and make the outcome of an attack with unmanned systems (including missiles) more uncertain.

Cyberwarfare

Autonomy is not a new development in the cyber realm. Automation is already a key component of any cyber-defence architecture. Anti-malware programs are designed to automatically identify and neutralize (known) malware. Cyberweapons generally need to operate autonomously—that is, outside direct human supervision—at least during key parts of their mission.[30] This was the case, for instance, for the Stuxnet virus.[31] However, recent advances in machine learning are changing the way that this automation or autonomy works as it changes the way in which cyberwarfare tools are designed and operated—whether for defensive or offensive purposes.

On the defensive side, machine learning methods have opened the possibility to spot new (i.e. unknown) types of malware and to detect suspicious activities in a network.[32] On the offensive side, machine learning facilitates the identification of zero-day vulnerabilities in an opponent's software systems. Machine learning in a nuclear context is a double-edge sword: it can both boost the protection of nuclear command-and-control infrastructure against cyberattacks and boost the enemy's capacity for cyberattacks against that infrastructure. Machine learning could enable a so-called left-of-launch operation: a cyber-offensive operation that would defeat the threat of a nuclear ballistic missile before it is launched.[33]

[27] Berlin, M. and Young, M., 'Automatic target recognition systems', *Technology Today*, no. 1 (2018), pp. 10–13.

[28] Hipple, M., 'Bring on the countermeasure drones', *Proceedings* (US Naval Institute), Feb. 2014.

[29] Scharre, P., *Robotics on the Battlefield*, part II, *The Coming Swarm* (Center for New American Security: Washington, DC, Oct. 2014).

[30] Guarino, A., 'Autonomous intelligent agents in cyber offence', eds K. Podins, J. Stinissen and M. Maybaum, *2013 5th International Conference on Cyber Conflict*, Proceedings, Tallinn, 4–7 June 2013 (NATO Cooperative Cyber Defence Centre of Excellence: Tallinn, 2013), pp. 377–89.

[31] Kile, S. N., 'Nuclear arms control and non-proliferation', *SIPRI Yearbook 2011: Armaments, Disarmament and International Security* (Oxford University Press: Oxford, 2011), pp. 363–87, p. 384.

[32] Polyakov, A., 'Machine learning for cybersecurity 101', Towards Data Science, 4 Oct. 2018.

[33] Ellison, R., 'Left of launch', Missile Defence Advocacy Alliance, 16 Mar. 2015.

Electronic warfare

Machine learning can bring major improvements to the field of electronic warfare in the same ways as for cyberwarfare.

On the defensive side, machine learning enhances anti-jamming capabilities as it opens the possibility to automate analysis and defence against new enemy signals.[34] In 2016 the US Defense Advanced Research Projects Agency (DARPA) launched a public challenge to develop systems with the capability to identify and analyse new enemy signals on the fly—that is, during the operation of the systems rather than afterward as is currently the case.[35]

On the offensive side, machine learning can be used to develop new jamming tools that could also play a role in a left-of-launch operation.

Physical security

Nuclear-armed states could combine the advances in machine learning and autonomy to automate the protection of their nuclear forces against physical attacks by terrorist groups or special forces. Autonomous robots—whether land, aerial or maritime—trained by machine learning are well suited for dull surveillance missions. Machine learning gives robots advanced detection capabilities while autonomy guarantees that they can keep a sharp and unblinking eye on the perimeters under protection. These systems could also be armed. Armed automated surveillance systems have, in fact, already been developed for border and perimeter protection. The most frequently discussed system is the robotic sentry weapon Super aEgis II, produced by the South Korean company DoDaam. The Super aEgis II is a gun turret equipped with sensors and an ATR system that can automatically detect, track and (potentially) attack targets—the system is designed to operate under human control, but it includes a 'fully autonomous' mode.[36]

It is debatable whether it would be operationally appropriate and, more importantly, lawful to use a robotic sentry weapon in fully autonomous mode to protect nuclear weapon systems (be it command and control or the nuclear weapons themselves). Some have argued that the limited ability of existing systems to distinguish between civilian and military targets and make a proportionality assessment would make their use in full autonomy unlawful.[37] Others have argued that the legality of such systems depends on the context and that using the systems without a human in the loop would not be problematic as long as it is deployed on

[34] Freeberg, S. J., 'Jammer not terminators: DARPA & the future of robotics', Breaking Defense, 2 May 2016.

[35] US Defense Advanced Research Projects Agency (DARPA), 'New DARPA Grand Challenge to focus on spectrum collaboration', 23 Mar. 2016.

[36] Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017), pp. 44–46.

[37] Brehm, M., *Defending the Boundary: Constraints and Requirements on the Use of Autonomous Weapon Systems under International Humanitarian Law and Human Rights Law*, Academic Briefing no 9 (Geneva Academy: Geneva, May 2017).

a perimeter where (*a*) it is reasonable to assume that there is no civilian presence and (*b*) circumstances would make the use of force proportionate.[38] It is safe to assume that the nuclear-armed states might have different perspectives on that issue.

Information warfare

One final area where the impact of machine learning and—to a lesser extent—autonomy could have strategic impact is information warfare. Machine learning offers new tools to directly or indirectly manipulate nuclear decision makers.

An example of direct use would be using GANs to create lifelike fake orders—in audio or video—that would trick nuclear weapon operators into launching a nuclear weapon or not responding to an attack. Higher command-and-control decision makers could also be indirectly tricked into doing or not doing something if their normal sources of information were tainted with fake information or fake opinion from people who would normal seem to be sensible.[39]

Should a nuclear-armed state decide to use machine learning algorithms for collection and processing of ISR information, this would open the possibility for an opponent to use a method known as data poisoning to undermine or manipulate the performance of early-warning systems.

V. Conclusions

Advances in machine learning and autonomy could be beneficial to all the key areas of the nuclear deterrence architecture: early warning and ISR; command and control; nuclear weapon delivery and non-nuclear counterforce operations (air defence, cybersecurity and physical protection of nuclear assets). The nature and magnitude of the impact will be different from one area to another. In some, the adoption of machine learning and autonomy could be transformative—that is, it may lead to notable operational and doctrinal changes; in other areas, it will merely lead to major qualitative improvements. None of the technologies presented above—even the most transformative—seems to have reached the stage where it could lead to a revolution in nuclear strategy. There are three reasons for this.

First are the safety and reliability problems deriving from the immaturity of the technology. Machine learning systems and autonomous systems still have numerous technical limitations that make their adoption risky from a command-and-control perspective. In the case of machine learning, the main source of concern is the uncertainty around the predictability and reliability of the systems caused by the algorithms' lack of transparency and explainability. For autonomous systems, it is the brittleness and the vulnerability to spoofing and cyberattacks

[38] Schmitt, M. N., 'Autonomous weapon systems and international humanitarian law: a reply to the critics', *Harvard National Security Journal*, Feature, 2013.

[39] On this scenario see chapter 13 in this volume.

that are of concern.[40] States would have to solve difficult testing and verification problems associated with the design of these systems to gain the confidence that they can be used safely in nuclear-related missions.

Second, even if these problems were to be overcome in the short term, the technology might still not be advanced enough to create a situation where nuclear-armed states could credibility threaten the survivability of each other's nuclear second-strike capability. Further advances in AI would be needed, and these would need to be supported by major progress in other enabling technologies, notably sensor and power technologies.

Third, capabilities offered by machine learning and autonomous systems could be offset or nullified by countermeasures. For instance, to counter advances made by an opponent in the field of ISR using machine learning and autonomy, a state might decide to exploit weaknesses in these technologies to its advantage. In the case of machine learning, that could involve resorting to data poisoning to deceive the enemy while, in the case of autonomous systems, it could be spoofing the sensors or jamming the communications network.

Thus, while advances in machine learning and autonomy will certainly bring a notable evolution in the conduct of the nuclear enterprise, they will not revolutionize the foundations of nuclear strategy. That being said, their adoption may still have a palpable impact on strategic relations and the power balance.[41]

[40] On these vulnerabilities see chapters 2–4 in part I of this volume.

[41] On the impact that the adoption of these technologies could have on strategic stability see chapters 9–14 in part III of this volume.

7. Artificial intelligence and the modernization of US nuclear forces

PAGE O. STOUTLAND

Recent years have seen dramatic advances in machine learning—an approach to artificial intelligence (AI) engineering that provides systems with the ability to automatically learn and improve from experience without being explicitly programmed.[1] These advances have enabled improvements in autonomous vehicles, image recognition, automated language translation, face recognition, online fraud detection and many other areas.[2] The military and security communities have begun to embrace these technologies, which bring promises of greater performance but also create new technical, legal and ethical challenges.[3]

For example, autonomous weapon systems enabled by machine learning may prove to be more capable of time-critical applications such as air defence, and machine learning may make possible new militarily relevant concepts such as swarms of drones. Numerous authors, however, have highlighted the legal and ethical implications of lethal autonomous weapon systems (LAWS) in which humans are not directly involved in making decisions about the use of lethal force.[4] There are also technical challenges: machine learning and, more broadly, AI-based approaches are currently fragile—multiple examples exist where, even without access to the software (so-called black box attacks), the systems may be spoofed, often with dramatic consequences.[5]

Recognizing the benefits, but also the pitfalls, of current machine learning approaches, this essay provides a framework and some initial thoughts on the implications of machine learning for the nuclear weapon force of the United States and the upcoming modernization. After a review of current modernization plans for US nuclear forces (section I), it describes the application of machine learning in nuclear weapons and related systems (section II).

I. US nuclear forces and modernization

US nuclear forces currently comprise three main pillars (the triad): (*a*) nuclear weapons carried by intercontinental ballistic missiles (ICBMs), (*b*) submarine-launched ballistic missiles (SLBMs), and (*c*) bombs and missiles

[1] Expert System, 'What is machine learning? A definition', [n.d.]. See also the presentation of machine learning in chapter 2 in this volume.

[2] See e.g. Future of Life Institute, 'Benefits & risks of artificial intelligence', [n.d.].

[3] Scharre, P. and Horowitz, M. C., *Artificial Intelligence: What Every Policymaker Needs to Know* (Center for New American Security: Washington DC, June 2018).

[4] Bostrom, N. and Yudkowsky, E., 'The ethics of artificial intelligence', eds K. Frankish and W. M. Ramsey, The Cambridge Handbook of Artificial Intelligence (Cambridge University Press: Cambridge, 2014), pp. 316–34.

[5] Brundage, M. et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation* (Future of Humanity Institute et al.: Oxford, Feb. 2018).

delivered by aircraft.[6] In addition to the weapons and delivery vehicles, US nuclear forces rely on early-warning systems and a command, control and communications (C3) system. To provide early indication of a nuclear attack, warning is provided by a combination of land-based radars and space-based satellites with infrared detection capabilities. Robust command, control and communications are provided by a mix of redundant communications assets that are able to transmit information from the early-warning systems, between decision makers and military officers, and, ultimately, to the weapons themselves.

The USA has approximately 1600 deployed strategic nuclear weapons.[7] The land-based ICBMs are kept in a 'prompt-launch' status, able to launch within minutes of being given the command. Similarly, submarines carrying nuclear weapons are able to launch in a slightly longer time period, perhaps of the order of 30 minutes.

While the current US nuclear forces have seen some level of continual upgrades, the existing assets have been in operation for decades. For example, the existing ICBMs (Minuteman III) first became operational in the 1970s and the SLBMs (Trident II) in 1990. Notably, one of the delivery aircraft for bombs, the B-52, first became operational in 1952. Because of the age of the force, an extensive modernization process has begun.

The overall modernization programme includes delivery systems, warheads, the production complex and the early-warning and command-and-control systems. This will include complete rebuilds of the missiles and new submarines and bombers. Warheads will be refurbished through life-extension programmes in which key components will be replaced. The total cost has been estimated at over US$1 trillion over 30 years, with over $40 billion to be spent on command and control and early warning alone.[8]

II. Machine learning in nuclear weapons and related systems

While many technical details will be determined in the coming years, modernization of the US nuclear forces will undoubtedly include additional use of digital systems, as has been the case for other civilian and military systems.[9] As use of digital systems and the accumulation of vast quantities of data expand, so do the potential applications of machine learning.

In nuclear weapon systems, the concept of automation—with autonomous systems being simple and predictive and largely based on if–then logical

[6] US Office of the Deputy Assistant Secretary of Defense for Nuclear Matters, *Nuclear Matters Handbook 2016* (Department of Defense: Washington, DC, 2016); and Congressional Research Service (CRS), *U.S. Strategic Nuclear Forces: Background, Developments and Issues*, CRS Report for Congress RL33640 (US Congress, CRS: Washington, DC, 21 Nov. 2018).

[7] In Jan. 2018 the USA had 1750 deployed warheads: 1600 strategic and 150 tactical. Kristensen, H. M., 'US nuclear forces', *SIPRI Yearbook 2018: Armaments, Disarmament and International Security* (Oxford University Press: Oxford, 2018), pp. 237–43, p. 237.

[8] Arms Control Association, 'U.S. nuclear modernization programs', Aug. 2018.

[9] Hagel, C., US Secretary of Defense, 'Reagan National Defense Forum keynote', 15 Nov. 2014; and McLeary, P., 'The Pentagon's third offset may be dead, but no one knows what comes next', *Foreign Policy*, 18 Dec. 2017.

frameworks, in contrast to AI or machine learning—is not new.[10] For example, while not based on machine learning technology, the Soviet Union's Dead Hand system in which, under certain conditions, nuclear missiles would be launched without a human in the loop is a known example of nuclear weapon automation.[11] However, US officials have been quite clear that the USA does not envision use of nuclear weapons without human authorization.[12]

Even if nuclear use decisions continue to be made by the US president in consultation with other officials, the use of machine learning is likely to expand during modernization. In the context of the nuclear weapon systems described in section I, potential applications of machine learning include warning systems, guidance systems, physical security, securing of communications systems (cybersecurity) and others. The following are some of the specific examples.

1. *Warning systems*. The land-based component of the US early-warning system is comprised of large phased-array radars.[13] Radars of this type depend on sophisticated processing algorithms that discriminate between the objects being searched for (i.e. nuclear missiles) and other objects that may be present (e.g. aeroplanes, birds etc.). Recent publications have highlighted the potential for machine learning-based algorithms to provide better discrimination abilities in radar applications.[14] If used in early-warning systems, this could in principle result in fewer false alarms.[15]

2. *Guidance systems*. Cruise missiles, including those carrying nuclear weapons, rely on sophisticated 'terrain-hugging' capabilities that allow them to fly close to the ground but not collide with mountains or tall buildings. While they rely on a range of technologies, recently the potential benefits of using machine learning to aid in navigation and targeting has been discussed by experts.[16]

3. *Cybersecurity*. In the cybersecurity area, machine learning is finding extensive use in the detection of malware.[17] While technical details of the US nuclear command, control and communications (NC3) system are not publicly available, it is not unreasonable to expect that network security would use sophisticated cybersecurity approaches, including machine learning.

[10] This definition of autonomous systems was put forward by Scharre, P., *Army of None: Autonomous Weapons and the Future of War* (W. W. Norton & Co.: New York, 2018). On different interpretations of automation and autonomy see also chapter 2 in this volume.

[11] Hoffman, D. E., *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy* (Anchor Books: New York, 2009); and Thompson, N., 'Inside the apocalyptic Soviet doomsday machine', *Wired*, 21 Sep. 2009. On the Dead Hand system see also chapters 5 and 8 in this volume.

[12] E.g. Clark, C., 'STRATCOM's Hyten on B-21, Columbia Class, NC3', Breaking Defense, 16 Apr. 2018.

[13] US Missile Defense Agency, 'Upgraded early warning radars, AN/FPS-132', Fact sheet, 23 July 2014.

[14] Rosa, I. M. D. et al., 'Classification success of six machine learning algorithms in radar ornithology', *Ibis*, vol. 158, no. 1 (Jan. 2016), pp. 28–42.

[15] Barrett, A. M., 'False alarms, true dangers? Current and future risks of inadvertent U.S.–Russian nuclear war', Rand Corporation, 2016.

[16] E.g. Rajagopalan, A., Faruqi, F. A. and Nandagopal, D., 'Intelligent missile guidance using artificial neural networks', *Artificial Intelligence Research*, vol. 4, no. 1 (Apr. 2015), pp. 60–76.

[17] Giles, M., 'AI for cybersecurity is a hot new thing—and a dangerous gamble', *MIT Technology Review,* 11 Aug. 2018. On the cybersecurity potential of machine learning see chapter 6 in this volume.

The above examples highlight how machine learning might be used in a modernized nuclear force. The potential performance benefits are likely to be significant and, despite the pitfalls cited below, may prove irresistible to developers and government sponsors.

While machine learning-based systems have enormous potential, it is often not fully understood how such systems work and it is not yet possible to characterize their robustness. In addition, experts have highlighted numerous failure modes, including the well-known examples in which slight modifications to an object (e.g. a roadway stop sign) have led an algorithm-based system to recognize it as something entirely different.[18] Others have highlighted the risk of unintended and harmful behaviour in machine learning systems (e.g. a cleaning robot disabling its vision so that it does not see the dirt to be cleaned).[19]

As a cautionary tale regarding the future implications of machine learning in nuclear weapon systems, it may be useful to consider the growth of digital technology. For example, advances in computing have brought a plethora of benefits that could not have been imagined 40 years ago. Unfortunately, early developers did not envision the unintended consequences (e.g. cybersecurity challenges) that would result from a pervasively networked society, and in the current situation critical systems have been implemented whose security cannot be ensured. Such a situation involving machine learning and nuclear weapons could be catastrophic.

III. Conclusions

First, a bit of good news: there appears to be agreement that any decision to use nuclear weapons should be made by a human. While technologies that increase the pace of warfare may at some point call this into question, at least for now there seems to be agreement on this in the USA. Other countries may see this differently—such as those with less secure second-strike capabilities, which may perceive that they could gain a competitive advantage by implementing greater degrees of automation.[20]

However, there will be strong motivations to include machine learning-based algorithms within the nuclear weapon systems themselves, in a range of platforms. Indeed, there may be genuine benefits that should not be ignored. For example, improved processing algorithms in early-warning radars could provide enhanced discrimination, thus lowering the possibility of a miscalculated use of a nuclear weapon. But since machine learning remains fragile, particularly when used in high-consequence systems, any decision to use it must be based on a careful consideration of the benefits and risks, including the potential for unintended behaviour or successful adversarial attacks.

[18] Szegedy, C. et al., 'Intriguing properties of neural networks', arXiv, 1312.6199, version 4, 19 Feb. 2014.

[19] Amodei, D. et al., 'Concrete problems in AI safety', arXiv, 1606.06565, version 2, 25 July 2016.

[20] Sullivan, T., 'NTI Seminar: A stable nuclear future? Autonomous systems, artificial intelligence and strategic stability with UPenn's Michael C. Horowitz', Nuclear Threat Initiative (NTI), 15 Nov. 2018.

Second, the broader nuclear policy implications of machine learning in nuclear systems must be considered. Even if machine learning in warning systems provides for better discrimination, would officials be comfortable making a decision based on a system for which there is no simple description of the way in which it 'decides' or even how it will perform in certain cases? Furthermore, how will the USA respond to an adversary's use of machine learning in its nuclear systems and what would be the implications for US nuclear forces and posture?

In sum, consideration of the implications, including the challenges and unknowns, of machine learning should not be deferred until such systems are operational. The implications of getting it wrong are too important. Rather, there must be careful analysis to fully understand the benefits, but also the unintended consequences, including how such systems might fail.

8. Autonomy in Russian nuclear forces

PETR TOPYCHKANOV

The Soviet Union and then the Russian Federation have been reluctant to provide a central role for autonomy in their nuclear weapons and related systems. Autonomy has been widely used in their nuclear command-and-control, ballistic missile defence, early-warning and now strike capabilities. But it has never replaced the human in the loop. The prospects for wider use of autonomy in these systems will depend not only on technological developments, but also on changes in Russia's nuclear posture and military planning.

This essay offers a brief analysis of the development of autonomy in the nuclear weapon and related systems of the USSR and Russia. For the current purposes, these nuclear force systems are understood to be the systems that compose a country's nuclear deterrence capabilities, that is, systems that are intended to deter an adversary from using conventional and nuclear force against that country. These systems include nuclear strike capabilities; missile and space defences (with both conventional and nuclear interceptors); intelligence, surveillance, target acquisition and reconnaissance (ISTAR) systems; early-warning systems; command-and-control systems; protection systems for nuclear forces and nuclear force-related facilities; training facilities; and equipment for assessment of the reliability and health of the military personnel working with nuclear force systems. Each of these systems is a promising area for application of autonomy.

This essay starts by reviewing the use of autonomy and automation in Soviet weapon development during the cold war (section I). It then looks at post-cold war developments in Russia (section II).

I. Nuclear weapon developments during the cold war

On 29 August 1949 the USSR conducted its first nuclear test to become the second nuclear-armed state (after the United States). Simultaneously, it started to develop its nuclear forces. This gave impetus to research work in computer science in the 1950s, in particular autonomy technologies. The fruits of this research were used first in missile defence and then in command and control.

Autonomy in missile defence and early warning

In the USSR the advance of computer science was driven to a great extent by the needs of its nuclear forces. From the early 1960s, the main engine behind progress in computer science were the demands of ballistic missile defence (BMD) and the early-warning system.[1]

[1] Revich, Yu. V., [Information technology and missile defence], ed. Yu. V. Revich, *Istoriya informatsionnykh tekhnology v SSSR* (Knima: Moscow, 2016), p. 48 (in Russian).

Box 8.1. Computers in missile defence and early warning

In the early 1970s, General Ivan Anureev, a consultant to the General Staff of the Soviet Armed Forces, wrote a report on the characteristics of ballistic missiles and space weapons and the principles of defence against them.

According to Anureev, the computers in such systems do not just control missile defence and early warning. As they process radar data, they can eliminate random mistakes caused by receiver noise by correlating the results of numerous observations. By measuring angles and range from a target, they are also able to calculate the altitude, direction and speed of missiles and the time they will take to reach the target. The computers also convert operator commands into instructions for the radar stations, and renew and convert obtained target data into information suitable for display.

Anureev identified three levels on which information is processed by computers in radar systems.

1. *On the sensor level*. The data output rate of the sensor can be increased by optimizing the target scanning, detection and tracking processes.

2. *In the data-processing system*. Here the question is reduced to lowering the time spent on filtering, smoothing (averaging) and correlating (comparing) the received radar information.

3. *In the decision-making system*. At the time, computers were assuming greater and greater significance, making it possible to take a decision in microseconds on the sequence of hitting objects in a group of targets.

Source: Anureev, I. I., *Antimissile and Space Defense Weapons* (Joint Publication Research Service: Arlington, VA, 1972), p. 119.

It was clear in the 1950s that computers could be central in detecting and intercepting incoming warheads. Most BMD and early-warning operations were seen as fully automated: autonomy was a required part of the design of these systems. This was due to the speed requirements of BMD and early-warning, since a human could not compete with machines in processing the data (see box 8.1). According to some assessments, the combined budget that the USSR spent on research and development of BMD and early-warning systems from the 1950s to the early 1970s surpassed the combined budget of the Soviet missile and space programmes.[2] Computer development was a significant part of these costs.

However, the signing of the 1972 Soviet–US Anti-Ballistic Missile Treaty (ABM Treaty) allowed the USSR to change its priorities and spend more on its command-and-control systems (including the Perimetr system), early-warning systems and offensive nuclear weapons.[3]

Autonomy in early warning

In the USSR there were doubts about the survivability of the nuclear arsenal after a massive counterforce strike by the USA. For that reason, until the end of the cold war the main principle of the Soviet nuclear posture was 'launch on warning'.[4]

[2] Revich (note 1), p. 48.

[3] Soviet–US Treaty on the Limitation of Anti-Ballistic Missile Systems (ABM Treaty), signed 26 May 1972, entered into force 3 Oct. 1972, not in force from 13 June 2002, *United Nations Treaty Series*, vol. 944 (1974), pp. 13–17.

[4] Kokoshin, A. A., [Past and present strategic stability: theoretical and practical questions] (KRASAND: Moscow, 2009), p. 81 (in Russian).

Box 8.2. The 1983 Petrov incident

On 26 September 1983, at about 00.30, Lieutenant Colonel Stanislav Petrov (1939–2017), who was in charge of monitoring the Oko early-warning system, saw on his display that a US attack was apparently under way. The warning system was reporting that five missiles originating from the United States were heading toward the Soviet Union. Petrov's role was to analyse the available information to determine whether this was a false alarm or a real attack, and in the latter case to report immediately to his superior commander.[a]

Petrov later explained in an interview that 'The main computer wouldn't ask me [what to do]—it was made so that it wouldn't even ask. It was specially constructed in such a way that no one could affect the system's operations.'[b]

Petrov and his team cross-checked the intelligence information but could not determine with certainty that it was a false alarm. Petrov nonetheless decided to report the incident as a false alarm to its superiors. He reportedly trusted his gut instinct.

[a] Hoffman, D. E., *The Dead Hand: Reagan, Gorbachev and the Untold Story of The Cold War Arms Race* (Icon: London, 2010), pp. 6–11.
[b] Likhanov, D., [40 minutes before World War III], *Rossiiskaya Gazeta*, 1 Sep. 2017 (in Russian).

In other words, the Soviet leadership planned a nuclear strike against the USA immediately after receiving information from early-warning systems about US missile launches.

The launch-on-warning principle increased the role of early-warning systems. The USSR's integrated, multilayered early warning system was planned in 1972. Based on this concept, the USSR created a network of over-the-horizon radars and warning satellites.[5] The integrated early-warning system was designed to accumulate data from various sources in an automated way. A real combat situation would allow operators only a limited time to analyse data that was automatically received from the early-warning system, and almost no time for separate double-checking.

This system never achieved the full capacity that was projected in 1972. The capacity that was achieved was not seen by the Soviet commanders as fully reliable. On many occasions the early-warning system falsely reported incoming missiles.[6] One of best-known examples of a false alarm was a 1983 incident with the Oko (Eye) early-warning system. This system consists of two satellites, Oko-1 and Oko-2, which can determine a missile trajectory by tracking its hot plume (see box 8.2). During this incident the early-warning system indicated a massive missile attack on the USSR. The officer in charge of the Oko command centre—Lieutenant Colonel Stanislav Petrov—had no chance to double-check the data—his role was to interpret the data on screen and report it to his commanders. The decision to launching a nuclear response was beyond his responsibility. The Oko system was one of several sources of information about a possible attack on the USSR, and it later became clear that the 1983 Oko alert was not supported by other sources. Nevertheless, the risk of an incorrect human command based on a false alarm was high at that time.

[5] Podvig, P., 'History and the current status of the Russian early-warning system', *Science and Global Security*, vol. 10, no. 1 (2002), pp. 21–60, pp. 23–24.
[6] Kokoshin (note 4), p. 82.

After the collapse of the USSR, the early-warning system continued to deteriorate. This led the Russian military to downgrade its role in nuclear command and control.[7] The launch-on-warning principle ceased to be central to Russian nuclear plans.

Automation of command and control

As mentioned above, the ABM Treaty allowed the USSR to prioritize offensive nuclear weapons and command and control. The rationale behind this shift was to prevent the USA from planning a nuclear first strike against the USSR. The ABM Treaty permitted each side to defend just one region with a BMD system. The USSR decided to defend Moscow. In the absence of BMD, the only possible way to prevent a US first strike was for the USSR to develop second-strike capabilities that could survive a first strike. The USA and the USSR later agreed the principle of preventing a first strike against each other through obtaining second-strike capabilities. This principle was mentioned in the 1990 Joint Statement on Future Negotiations on Nuclear and Space Arms and Further Enhancing Strategic Stability as a key principle of 'enhancing strategic stability'.[8]

The survivability of the second-strike capability meant the survivability of not only offensive weapons, but also command-and-control systems. In 1974, two years after the ABM Treaty was signed, the Soviet Government decided to start the research and development of the highly automated Perimetr (Perimeter) command-and-control system.[9] The purpose of this system was to initiate a mass retaliation with all remaining means in case an adversary should hit Soviet territory with a first strike and the political and military leadership could not operate normally, whether because of disruption of communications or decapitation of the leadership. Perimetr became operational in 1985.

How the Perimetr system operates[10]

Perimetr may be alerted in two ways. In the first, it can be alerted by a human. The second way is for Perimetr to alerted itself because of data received that confirms a nuclear attack, based on information from land-, sea-, air- and space-based sensors. The system then requires a yes or no responses from the General Staff of the Armed Forces.

If the supreme commander (now the president of Russia) survives the first strike and is reachable, the General Staff addresses the request from the Perimetr system to him or her, and then forwards the decision to the Perimetr system. If Perimetr receives no response from the General Staff, it requires a yes or no response from

[7] Podvig (note 5), p. 54.

[8] Soviet–United States Joint Statement on Future Negotiations on Nuclear and Space Arms and Further Enhancing Strategic Stability, Washington, DC, 1 June 1990.

[9] Valagin, A., [Assured retaliation: how the Russian 'Perimetr' system works], *Rossiiskaya Gazeta*, 22 Jan. 2014 (in Russian).

[10] Elsewhere in this volume and in other publications in English, the name Dead Hand is used for the command system based on the Perimetr command rockets. However, the name Dead Hand has been never supported in Russian language open sources, and so the present author prefers not to use it.

the so-called nuclear briefcase 'Cheget', which is part of the 'Kazbek' command-and-control system of the Strategic Rocket Forces.[11] If there is no response from the nuclear briefcase, Perimetr requests a yes or no response from any command centre of the Strategic Rocket Forces. Only after receiving no response from any of these sources is Perimetr designed to initiate retaliation.

It is difficult to imagine Perimetr being alerted in the absence of a nuclear attack or when an adversary is using only conventional and cyber means, because the key precondition for it being alerted is to receive data from sensors that confirm that a nuclear attack has happened.

Thinking the unthinkable, if Perimetr were to be alerted in the absence of an actual nuclear strike against the country, the system is designed in such a way that all changes of status are transparent to the authorized political and military commanders and all these changes may be cancelled by authorized humans at any stage. The principle of having a human in the loop remains the basis for Perimetr, while at the same time allowing for fully automated operation.

There are two views on the capacity of the Perimetr system. According to one view—shared by the authors of the book *Russian Strategic Nuclear Forces*—the full functionality was never activated, and the system only operated in the form of command rockets.[12] Another view is presented by Colonel Valery Yarynich, a former officer of the Strategic Rocket Forces and later the General Staff. In an interview in 2009 he described the full operational capacity of the Perimetr system and stated that it was 'continuously being upgrade'.[13]

When it became no longer relevant after the end of the cold war and the change in the relations between Russia and the West, the Perimetr system was frozen in 1995 as part of the de-alerting of nuclear forces.[14] This was a unilateral decision made by Russia without similar de-alerting steps by the USA. While the system was made non-operational, it was not dismantled. However, over time, part of the infrastructure became outdated.

II. Post-cold war developments in Russia

Change of Russia's security environment: Perimetr operational again

In 2011 the commander-in-chief of the Russian Strategic Rocket Forces, Sergei Karakayev, confirmed that the Perimetr system had become operational again.[15] It was not simply a resumption of the previous system: it went through several

[11] Arbatov, A., 'Democracy and nuclear weapons', *Russia in Global Affairs*, 30 July 2005.

[12] Podvig, P. (ed.), *Russian Strategic Nuclear Forces* (MIT Press: Cambridge, MA, 2001), pp. 65–66.

[13] Thompson, N., 'Inside the apocalyptic Soviet doomsday machine', *Wired*, 21 Sep. 2009.

[14] Shirokorad, A., ['Dead Hand' is more dangerous than 'Aegis' and 'Tomahawk'], *Nezavisimoye Voyennoe Obozreniye*, 9 Apr. 2010 (in Russian).

[15] Baranets, V., [Commander-in-Chief of the RSRF Lieutenant General Sergei Karakayev: 'Vladimir Vladimirovich was right, we can destroy the USA in less than a half of hour'], *Komsomol'skaya Pravda*, 16 Dec. 2011 (in Russian).

rounds of modernization.[16] The reason for reviving Perimetr has never been explained by Russian officials.

Two recent statements about Perimetr may indicate new developments in Russia's nuclear planning. These developments suggest some changes in the role of automation in Russia's command-and-control and offensive nuclear systems.

The first of these statements was made by President Vladimir Putin on 18 October 2018: 'any aggressor should know that retaliation is inevitable and they will be annihilated. And we as the victims of an aggression, we as martyrs would go to paradise while they will simply perish because they won't even have time to repent their sins.'[17]

The second statement was made on 8 November 2018 by Colonel General Victor Esin, former chief of staff and vice-commander-in-chief of the Russian Strategic Rocket Forces, who confirmed that the Perimetr system is operational and upgraded.[18] At the same time, he said that the system will not be effective if the USA withdraws from the 1987 Treaty on the Elimination of Intermediate-Range and Shorter-Range Missiles (INF Treaty) and deploys the ground-launched missiles in Europe that are currently banned.[19] If this happens, the USA will be able to turn against Russia long-, intermediate- and short-range delivery vehicles for nuclear weapons and high-precision conventional weapons of various ranges. As a result, Russia will not have many second-strike capabilities after a possible first strike from the USA.[20] This possibility undermines the main purpose of Perimetr, which is to initiate a mass retaliation with all the remaining means.

Both statements confirm that the Perimetr system is operational and has the same purpose that it was designed for in the Soviet era. However, unlike during the cold war, the current size of the Russian nuclear arsenal throws this purpose into question. When the Perimetr first became operational in 1985, the USSR had 39 197 nuclear warheads.[21] As of 2018, the total Russian nuclear arsenal consists of an estimated 6850 warheads, with 1420 of them being operationally deployed.[22] If Russia were to be attacked by the USA, only part of this arsenal would survive. This part would not be enough for a mass retaliation, especially taking into account that the first strike may include the nuclear, conventional, cyber and electronic

[16] Valagin (note 9).

[17] President of Russia, 'Meeting of the Valdai International Discussion Club', 18 Oct. 2018.

[18] Odnokolenko, O., [Colonel General Victor Esin: 'If the Americans finally deploy their missiles in Europe, we will have to replace the launch under attack doctrine with the doctrine of pre-emptive strike'], *Zvezda*, 8 Nov. 2018 (in Russian).

[19] Soviet–US Treaty on the Elimination of Intermediate-Range and Shorter-Range Missiles (INF Treaty), signed 8 Dec. 1987, entered into force 1 June 1988, *United Nations Treaty Series*, vol. 1657 (1991), pp. 4–167.

[20] Odnokolenko (note 18).

[21] Norris, R. S. and Kristensen, H. M., 'Global nuclear weapons inventories, 1945–2010', *Bulletin of the Atomic Scientists*, vol. 66, no. 4 (July–Aug. 2010), pp. 77–83, p. 81.

[22] Kristensen, H. M., 'Russian nuclear forces', *SIPRI Yearbook 2018: Armaments, Disarmament and International Security* (Oxford University Press: Oxford, 2018), pp. 244–51, p. 247; and US Department of State, Bureau of Arms Control, Verification and Compliance, 'New START Treaty aggregate numbers of strategic offensive arms', Fact sheet, 1 Sep. 2018.

means of the USA and its allies. A Russian retaliation may also be limited by the USA's growing BMD capabilities.

Esin's argument was about possible consequences for Russia of the end of the INF Treaty, from which the USA is expected to withdraw on 2 August 2019.[23] If ground-based missiles of intermediate and shorter ranges, deployed in Europe, are added to the US capabilities, then Russia's defensive nuclear posture will become less effective in providing nuclear deterrence against the USA.

Automation in offensive and defensive nuclear postures

The risks outlined above have made the Russian authorities contemplate replacing the defensive nuclear posture with an offensive one. The Perimetr system is a central part of the former, but it will not play the same role in the latter. An offensive nuclear posture is based on the possibility of a pre-emptive nuclear strike. The purpose of this strike is to prevent a nuclear attack or stop a conventional attack from an adversary. At the same time, a pre-emptive strike should not provoke a mass nuclear retaliation from the adversary.

An offensive nuclear posture would thus reduce the role of the highly automated Perimetr system. Such a posture would require carefully calibrated plans for limited use of nuclear weapons. These plans would be changeable due to Russia's highly volatile security environment.

The defensive nuclear posture has highly automated command and control as a possible substitute for a human-based system in the cases when the political and military leadership cannot operate normally. The offensive posture would not need this level of command-and-control automation at the beginning of the armed conflict.

If the conflict continues with retaliation by the adversary, then Perimetr may be used for the response. But it clearly would not be a mass use of nuclear weapons because part of the Russian arsenal would have been used in a pre-emptive strike and another part would have been eliminated by the adversary's retaliation. In the final stage of an armed conflict started with a nuclear pre-emption, the use of Perimetr-like command-and-control systems seems to have no political or military sense.

However, an offensive nuclear posture may create a demand for new capabilities for the limited use of nuclear weapons. One of these was presented by President Putin in 2018: the Poseidon nuclear-powered unmanned underwater vehicle (UUV).[24] Poseidon (also known as Status-6) is an autonomous system that, when and if it is commissioned, should operate according to commands from command centres. Poseidon is to be launched from a nuclear-powered submarine. It may perform various missions depending on the type of payload it is carrying, which could be a nuclear warhead or surveillance equipment. Its mission should theoretically be under human supervision from start to end with use of trailing

[23] US Department of State, 'U.S. intent to withdraw from the INF Treaty', Press statement, 2 Feb. 2019.

[24] President of Russia, 'Presidential address to the Federal Assembly', 1 Mar. 2018. The Poseidon system is also discussed in chapters 6, 9, 11 and 14 in this volume.

wire antennas.[25] However, the key question is how to secure such a level of control of this nuclear-powered UUV without risking the loss of the connection to it and compromising the secrecy of its movements, especially in a combat situation.

The underwater trials of the Poseidon engines started in late 2018. The system is scheduled to be commissioned before the end of the State Armament Programme for 2018–27.[26]

III. Conclusions

The nuclear posture of Russia remains defensive. Autonomy is central to the Perimetr command-and-control system for a retaliatory attack and the early-warning systems. Based on publicly available information, the former has been upgraded by Russia while the latter has degraded since the collapse of the USSR.

The recent developments in Russian–US relations may bring serious changes in Russia's nuclear posture and military planning. The posture may become offensive and the planning may include options for the limited use of nuclear weapons.

The role of autonomy in the nuclear weapons and related systems will reflect these changes. There may be a shift from the central role of highly automated early-warning and retaliatory capabilities to the wider use of autonomous strike capabilities.

[25] Ramm, A., Kornev, D. and Boltenkov, D., [Leak put under a microscope], *Voenno-Promyshlennyi Kurer*, 25 Nov. 2015 (in Russian).

[26] ' [Underwater tests of nuclear propulsion system of unmanned vehicle 'Poseidon' have been started], Oruzhiye Rossii, 28 Dec. 2018 (in Russian).

Part III. Artificial intelligence, strategic stability and nuclear risk: Euro-Atlantic perspectives

How and to what extent could the current status quo between the nuclear-armed states be undermined by their adoption of systems based on artificial intelligence (AI), be it for conventional or nuclear weapons? This is the question that scholars from both sides of the Atlantic address in the following six essays.

The first three essays—by Michael Horowitz of the United States (chapter 9), Frank Sauer of Germany (chapter 10) and Jean-Marc Rickli of Switzerland (chapter 11)—provide a general overview of how military applications of AI could alter the foundations of strategic stability in the Euro-Atlantic context.

The following two essays—by Justin Bronk of the United Kingdom (chapter 12) and Shahar Avin of Israel and S. M. Amadae of the USA (chapter 13) focus on the risks deriving from the application of AI in two specific technology areas: unmanned combat aerial vehicles (UCAVs) and cyberwarfare, respectively.

In the final essay (chapter 14), two high-level United Nations practitioners, Anja Kaspersen and Chris King, share their personal views on how the risks posed by the military application of AI in the field of nuclear weapons and doctrines could be dealt with by the international community.

VINCENT BOULANIN

9. Artificial intelligence and nuclear stability

MICHAEL C. HOROWITZ

The question of how advances in artificial intelligence (AI) could influence the probability of nuclear war represents one of many important questions surrounding how AI developments may shape the international security environment. Despite the fear of 'killer robots' in the media, most uses of AI will involve image recognition, data analysis and other related systems, rather than battlefield weapons. Even so, these applications could significantly influence nuclear stability in some cases.

In general, how nuclear-armed states think about using autonomous systems may depend most on the extent to which they view their second-strike capabilities as vulnerable. The more vulnerable they view these capabilities to be, the more likely they are to integrate autonomous systems, especially those that may speed up decision-making or cut the human out of the loop. Fundamentally, insecure nuclear-armed states worry about decapitation, whether due to conventional or nuclear weapons. A key potential benefit of autonomous systems is the ability to make decisions more quickly—and autonomously. An insecure nuclear-armed state would therefore be more likely to automate nuclear early-warning systems, use unmanned nuclear delivery platforms or, due to fear of rapidly losing a conventional war, adopt nuclear launch postures that are more likely to lead to accidental nuclear use or deliberate escalation.[1]

For the purposes of this essay, the term artificial intelligence refers to narrow (or weak) applications of AI.[2] These are algorithms designed for a specific task, such as AlphaGo Zero (designed to play the game go), and which cannot innovate beyond their initial programming. The alternative to narrow AI is artificial general intelligence (AGI or strong AI), which has the ability to innovate independently and to write new programming to do new tasks. This essay focuses on narrow AI because advances in narrow AI are more certain to occur and because it is more intellectually tractable to analyse.

In the broad category of robotics and autonomous systems, AI represents something different from the remotely piloted robotic systems that many militaries operate today. Unmanned aerial vehicles (UAVs) such as the MQ-9 Reaper are still piloted, in the same way that an F-18 combat aircraft is piloted. The pilot is just not on board the aircraft. Even though some UAVs incorporate elements of autonomy to assist with take-off and landing, for example, those systems are much more akin to autopilot on commercial aircraft than anything else.

What does this have to do with nuclear weapons? Both more and less than many commentators assume. This essay considers the intersection of AI and the

[1] The term 'unmanned' is used here for consistency with the rest of this volume. A better, ungendered term would be 'uninhabited'.

[2] On the definition of artificial intelligence and the distinction between narrow (or weak) and general (or strong) AI see chapter 2 in this volume.

nuclear weapon complex across three categories: nuclear command and control (in section I), unmanned nuclear delivery platforms (in section II), and the impact of conventional military uses of autonomous systems on the potential for nuclear escalation (in section III).

I. AI and nuclear command and control

Excluding a first strike, the first step in the process leading up to the possible use of nuclear weapons is how a nuclear-armed state attempts to detect whether another country is launching nuclear weapons and how it responds. Many countries already automate parts of their nuclear weapon infrastructure, especially advanced nuclear powers such as the United States.[3] This includes early warning, command and control, and missile targeting. Advances in AI could lead to the expansion of the use of autonomous systems in command and control. For example, states could decide to automate additional components of early warning because autonomous systems can detect patterns and changes in patterns faster than humans. This could have potential benefits for nuclear security and stability, because well-functioning algorithms could give decision makers more time in a complex environment. Moreover, autonomous systems could represent another form of redundancy that helps to ensure the dissemination of launch orders in the worst case.

However, the 1983 Petrov incident illustrates a clear downside to fully automating command and control. In this case, the Soviet Oko satellite-based early-warning system reported a false alarm—the launch of five US intercontinental ballistic missiles (ICBMs). No missiles had been launched. Lieutenant Colonel Stanislav Petrov was the watch officer on duty. It was his job to alert Soviet leadership of a US attack. While the automated systems reported the 'highest' confidence that a missile strike was occurring, Petrov stated that he 'had a funny feeling in [his] gut'. He instead reported a system malfunction, rather than a nuclear strike.[4]

The risk is that a future incident could lead to escalation, instead of a malfunction report, for two reasons. First, a decision to fully automate early warning would mean that there was no human operator—no Petrov—to prevent a false alarm from escalating. To be fair, however, it seems unlikely that a country would cut humans entirely out of the early-warning process. Second, automation bias could mean that a future Petrov trusts the algorithm and instead reports that an attack is under way.[5] While also unlikely, academic research on automation bias suggests that this is a real risk.[6]

[3] Blair, B. G., *The Logic of Accidental Nuclear War* (Brookings Institution: Washington, DC, 1993).

[4] Aksenov, P., 'Stanislav Petrov: the man who may have saved the world', BBC, 26 Sep. 2013; and Hoffman, D., '"I had a funny feeling in my gut"', *Washington Post*, 10 Feb. 1999, p. A10.

[5] On automation bias and the Petrov incident see chapter 5 in this volume.

[6] Skitka, L. J., Mosier, K. L. and Burdick, M., 'Does automation bias decision-making?', *International Journal of Human–Computer Studies*, vol. 51, no. 5 (Nov. 1999), pp. 991–1006; and Cummings, M., 'Automation bias in intelligent time critical decision support systems', American Institute of Aeronautics and Astronautics 1st Intelligent Systems Technical Conference, Chicago, IL, 20–22 Sep. 2004.

Automation bias is when humans, lulled into a false sense of security by the repeated success of algorithms, stop questioning them and trust them completely. This generates cognitive offloading, where humans become unlikely to question an autonomous system even in a scenario where an unbiased human might recognize that an algorithm-based system is reporting incorrect information.[7]

Thus, while the introduction of autonomous systems into nuclear command-and-control offers potential benefits in terms of faster recognition of a strike, there are potential risks due to the potential for automation bias, even if there is still a human in the chain of command.

II. AI and nuclear delivery platforms

An autonomous nuclear delivery platform would be an autonomous version of a combat aircraft, bomber or submarine carrying an armed nuclear weapon. Why put nuclear weapons on a UAV? A country may fear the hacking of its communications lines in a crisis. Preprogrammed autonomous systems could potentially be invulnerable to such interference. In general, however, autonomous nuclear delivery platforms seem risky.

The risks associated with an autonomous platform with nuclear weapons, which would eliminate positive human control over nuclear weapon use, seem obvious. Hacking or spoofing could make a system vulnerable to capture or malfunction even before factoring in the chance that the brittle character of an algorithm leads to a malfunction.

It is countries that feel relatively insecure about their nuclear arsenals that should be the most likely to turn to unmanned nuclear delivery platforms. For example, the USA can feel relatively secure that it has the most advanced nuclear arsenal in the world. Thus, US military leaders, despite the well-known reluctance of the USA to rule out potential capabilities, have clearly stated their resistance to arming autonomous vehicles, or even remotely piloted vehicles, with nuclear weapons. For example, in 2016 General Robin Rand, head of Air Force Global Strike Command, said: 'We're planning on [the B-21 Raider long-range bomber] being manned . . . I like the man in the loop; the pilot, the woman in the loop, very much, . . . particularly as we do the dual-capable mission with the nuclear weapons.'[8]

In contrast, Russia is generally secure, but fears the USA's conventional military advantage and its nuclear arsenal. Thus, in 2012 Lieutenant General Anatoly Zhikharev, commander of Long-Range Aviation, stated that Russia could field an unmanned nuclear bomber by the 2040s.[9] Russia also leaked its discussions on building the Poseidon nuclear-armed unmanned underwater vehicle (UUV).[10]

[7] Cummings, M. L., 'Creating moral buffers in weapon control interface design', *IEEE Technology and Society Magazine*, vol. 23, no. 3 (fall 2004), pp. 28–33, pp. 29–30.

[8] Hodge Seck, H., 'Air Force wants to keep "man in the loop" with B-21 Raider', DefenseTech, 19 Sep. 2016.

[9] 'Russia could deploy unmanned bomber after 2040—Air Force', RIA Novosti, 2 Aug. 2012.

[10] Mizokami, K., 'Pentagon confirms Russia has a submarine nuke delivery drone', *Popular Mechanics*, 8 Dec. 2016. The Poseidon UUV is also discussed in chapter 6, 8, 11 and 14 in this volume.

The system (also known as Status-6) could submerge to 1000 metres and deploy off the coast of a potential adversary for an indefinite period. Since it would not have to return to Russia often (if ever) once deployed, the system would be extremely difficult for even US submarines to detect. It is Russia's relative insecurity in comparison to the USA (despite their broad nuclear advantage relative to any other country) that has arguably helped to drive its interest in these autonomous systems, even if they never become reality. The way that a much weaker, less secure, nuclear power might be more likely to consider autonomous platforms armed with weapons of mass destruction is further demonstrated by discussions by the Democratic People's Republic of Korea (DPRK, or North Korea) of potentially using UAVs to deliver chemical or radiological weapons against the Republic of Korea (South Korea).[11]

III. Conventional military uses of AI and nuclear escalation

Perhaps the greatest risk of nuclear escalation arising from the use of autonomous systems and AI may come from the way that conventional military uses of AI could place pressure on nuclear powers to adopt unstable launch postures or even to strike first in a crisis. One of the primary benefits of AI is the ability of machines to make judgments faster than people. Operating at machine speed could be an advantage on the battlefield, because countries using autonomous systems could outpace those with human operators.[12]

However, one county's ability to potentially win a conflict at machine speed means another country could also lose at machine speed. And the fear of losing at machine speed could encourage a weaker nuclear power—especially one that is not confident in its second-strike nuclear capabilities—to adopt nuclear use postures generally thought to be unstable, such as pre-delegating potential nuclear use early in a conflict or a launch-on-warning posture. In the worst case, the fear of losing quickly at the outset of a conflict could even lead to first-strike instability, as a country that fears decapitation decides to attack first, with nuclear weapons, in an attempt to avoid potential future defeat.

Note that there is nothing about the systems necessarily being autonomous that generates instability in this case. It is the increasing speed of warfare. Thus, other potential developments, such as hypersonic weapons, could also generate instability.[13] Finally, these postures could lead to accidents and miscalculation as countries once again put their nuclear arsenals on a hair trigger due to the fear of rapid decapitation.

The impact of combat at machine speed at the conventional level on the risk of nuclear escalation could be exacerbated by uncertainty about how autonomous

[11] Mizokami, K., 'Experts: North Korea may be developing a dirty bomb drone', *Popular Mechanics*, 28 Dec. 2016.

[12] Horowitz, M. C., 'When speed kills: autonomous weapon systems, deterrence, and stability', Working paper, University of Pennsylvania, Apr. 2019.

[13] Acton, J. M., 'Hypersonic weapons explainer', Carnegie Endowment for International Peace, 2 Apr. 2018.

systems will function on the battlefield. It is still early in—or even prior to—the age of advanced autonomous military systems, and countries are uncertain about how autonomous systems will function.

IV. Conclusions

Automation could have both positive and negative impacts on the risk of nuclear stability, and the negative impact may be highest for countries that doubt the efficacy of their second-strike capabilities. Autonomous systems offer potential advantages in terms of speed and reliability, which could make them tempting for use in early-warning systems. However, autonomous systems are also brittle—and when the operating environment changes, even in small ways, they become more likely to fail.

However, automation could improve safety and reliability in nuclear operations in some cases. Simple and repetitive tasks where human fatigue, anger and distraction could interfere are ripe for the use of algorithms. The optimal situations might be those where countries use humans and machines in combination—as long as training and doctrine help human operators hedge against automation bias.

Most importantly, it is conventional applications of AI that could, in the end, lead to the greatest increase in the risk of nuclear escalation. The ability of autonomous systems to increase the speed and accuracy of conventional war could place pressure on nuclear powers in a crisis situation, especially those less sure about their second-strike deterrent. Fearing decapitation at the outset of a conflict, these countries may be more likely to turn to dangerous nuclear-launch doctrines, such as launch on warning. In the worst (albeit unlikely) case, first-strike instability could result.

It is critical, however, to keep in mind the large degree of uncertainty surrounding advances in AI. It is possible, for example, that cyber vulnerabilities generate much more hesitancy about the use of autonomous systems than is described above. Regardless, the intersection of AI and nuclear weapons will be critical for the international security environment in the years ahead.

10. Military applications of artificial intelligence: Nuclear risk redux

FRANK SAUER

There is considerable hype surrounding the military use of artificial intelligence (AI) and the approach primarily responsible for current advances in the field: machine learning. With ongoing efforts to use them to increase autonomy in weapon systems, new risks arise, including possible detrimental effects on strategic stability. Moreover, AI and machine learning also provide new possibilities to manipulate the information landscape in which nuclear decision-making takes place. Consequently, nuclear risk-mitigation measures, such as the adoption of no-first-use doctrines or a lowered weapon alert status, are more pertinent than ever.

This essay starts by defining AI and machine learning and outlining the misconceptions that surround them (section I). It then looks at why they are being applied in conventional weapon systems and what the associated risks are (section II). The nuclear risks that can arise from the application of autonomy in both conventional and nuclear forces are then assessed (section III).

I. AI and machine learning

Defining AI and machine learning

AI is a broad concept, with no uniform definition, and the goal posts of what is considered to be artificially intelligent are constantly shifting—yesterday's monumental AI breakthrough is just ordinary software today (e.g. computers playing chess).[1] One key element is automation, rendering it an essential aspect of the most commonly used working definitions of AI. In fact, the increased potential for task automation—which is driven primarily by private investment and the civilian technology companies responsible for most of the innovation in AI—is at the heart of the current string of AI breakthroughs and successes, big and small. These range from the automated sorting of smartphone photos in the cloud to, in the hopefully not-too-distant future, self-driving cars. Emphasizing automation means defining AI as software-based techniques and procedures deployed to automate tasks for which the application of human intelligence was previously required.

Machine learning is the use of algorithms and advanced statistical models to improve task performance. In that regard, it is the approach primarily responsible for most of the recent AI advances. Machine learning, especially deep learning using neural networks, is now yielding impressive results due to increases in

[1] Levy, S., 'What Deep Blue tells us about AI in 2017', *Wired*, 23 May 2017. On the definitions of AI and machine learning see chapter 2 in this volume.

computational power and the availability of large volumes of labelled data on which the systems can be trained.[2] Machine learning in its deep learning variety is a powerful tool for building systems for pattern recognition, for example in still or moving images and in written or spoken text. It constantly spawns a variety of new and exciting civilian applications.

The misconceptions surrounding AI and machine learning

AI and machine learning are still simultaneously over- and underestimated by both the general public and policymakers. They are being overestimated because the 'intelligence' component of the term AI evokes the wrong association, namely with human learning and human intelligence. These both differ significantly from the nature of AI and machine learning and what they are currently capable of. After all, machine learning-based AI is limited to extremely narrow tasks. It is greedy (i.e. hungry for immense of amounts of data), brittle (i.e. failing spectacularly when confronted with a task that differs slightly from what it was trained for) and opaque (i.e. generating unexplainable outputs, essentially rendering it a black box that is impossible to debug).[3] In other words, AI and machine learning are not at all comparable to the flexible and generalized skills and problem-solving competence that come with human learning and intelligence.

The poor understanding of machine learning in general, in particular its stochastic rather than deterministic nature, and a misjudgement of its strengths and weaknesses are in large part responsible for why the current hype surrounding the military applications of AI and machine learning is so dangerous. The limits of the technology would suggest a slow and careful introduction of AI into the military. Instead, the notion of AI as an 'enabling technology' is adopted imprudently.[4] The result is the misguided hope that almost every aspect of the military can soon be enhanced by it in some shape or form.

At the same time, the potential risks remain underappreciated. Such risks include biased training data leading to detrimental effects in various contexts of algorithmic decision-making. In the civilian sphere such risks can touch on various aspects of life, such as tax collecting, loan granting and even medical diagnoses.[5] In the military sphere, where the highest premium is placed on the additional speed that can be gained from automation, the risks are also manifold—and some of them give rise to fundamental issues.

[2] Marcus, G., 'Deep learning: a critical appraisal', arXiv, 1801.00631, 2 Jan. 2018.

[3] Marcus (note 2).

[4] Horowitz, M. C., 'Artificial intelligence, international competition, and the balance of power', *Texas National Security Review*, vol. 1, no. 3 (May 2018), pp. 37–57.

[5] As an example of the latter see Chen, A., 'IBM's Watson gave unsafe recommendations for treating cancer', The Verge, 26 July 2018.

II. AI, machine learning and autonomy in weapon systems

The past, present and future of weapon autonomy

Militaries are exploring the use of AI and machine learning for a variety of purposes: from logistics via predictive maintenance to strategic foresight as well as in improved decision aids for command and control or battle management. However, automation yields its most immediate and crucial battlefield value in the reduction of the time required to complete the so-called targeting cycle—the process of finding, fixing, tracking, selecting and engaging a target and then assessing the outcome of the engagement.[6] The completion of this cycle can be sped up at any of its six stages.

For example, Project Maven is a cooperation between the United States Department of Defense (DOD) and the US multinational technology company Google in which Google provides automated analysis of video footage from unmanned aerial vehicles (UAVs) to reduce the workload of human analysts. It focuses on the first part of the targeting cycle: finding, fixing and tracking.[7]

When already engaged in battle, the biggest advantage is to be gained from speeding up the next part of the targeting cycle—selecting and engaging targets. The current scholarly and diplomatic debate around autonomy in weapon systems is mostly concerned with these two final, so-called critical functions.[8] For a weapon system to be fully autonomous as defined by the US DOD and the International Committee of the Red Cross (ICRC), it must 'once activated, . . . select and engage targets without further intervention by a human operator'.[9]

Weapon systems that autonomously engage targets are not new, of course, and they are not necessarily problematic. Systems capable of engaging targets without human intervention, such as the Phalanx close-in air-defence system on navy vessels, have been in use for decades.[10] However, such systems are stationary, perform only the same preprogrammed actions repeatedly, and usually direct their fire only at incoming ordnance. If that is the case, especially since human life is not threatened, then weapon autonomy of the 'old' kind is of little cause for concern. In contrast, the 'new' kind of weapon autonomy that is currently raising alarms is present in mobile systems that operate in dynamic, unstructured, open

[6] On the use of the targeting cycle as an analytical framework for weapon autonomy see the work of Merel Ekelhof, most recently Ekelhof, M. A. C., 'Lifting the fog of targeting: "autonomous weapons" and human control through the lens of military targeting', *Naval War College Review*, vol. 71, no. 3 (summer 2018), pp. 61–94. See also the reports of the International Panel on the Regulation of Autonomous Weapons (iPRAW), most recently iPRAW, *Concluding Report: Recommendations to the GGE* (Stiftung Wissenschaft und Politik: Berlin, Dec. 2018).

[7] The project is reportedly due to end in 2019 following protests by Google employees. Wakabayashi, D. and Scott, S., 'Google will not renew Pentagon contract that upset employees', *New York Times*, 1 June 2018. On Project Maven see also chapters 5, 6 and 11 in this volume.

[8] International Committee of the Red Cross (ICRC), *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, Expert meeting, Versoix, Switzerland, 15–16 Mar. 2016 (ICRC: Geneva, Aug. 2016).

[9] US Department of Defense, 'Autonomy in weapon systems', Directive no. 3000.09, 21 Nov. 2012, updated 8 May 2017, p. 13. See also International Committee of the Red Cross (note 8), p. 8.

[10] Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017), p. 38.

environments over longer periods of time, engaging a variety of targets, including inhabited targets or individual humans. Fully autonomous weapon systems meeting this definition have, albeit with narrow tasks, already been fielded. A prominent example is the Israeli anti-radar loitering munition Harpy; in this case, the scope of application is limited to cruising over an area and engaging enemy air-defence radar systems.[11]

Risks deriving from the advance of autonomy in weapon systems

Autonomy begets autonomy because speed—defined as the ability to complete the targeting cycle before an adversary does—promises a key tactical advantage. The risks deriving from this 'race for speed' towards full weapon autonomy touch international law and ethics as well as global security and stability.[12]

For instance, there are serious doubts about the compliance of autonomous weapon systems with requirements of international humanitarian law, especially the distinction between civilians and combatants or the proportionate use of military force.[13] Moreover, the notion of delegating the legally required human judgment to a machine is ethically questionable in itself, regardless of the machine's performance. After all, the guiding principle of respect for human dignity dictates that machines should generally not be making life-or-death decisions.[14] With regard to global stability, the unpredictable behaviour of autonomous weapon systems in scenarios where multiple algorithmically controlled weapon systems would come to interact is of special concern.[15] It has sparked worries about unwanted, split-second military escalations, triggered and cascading too fast for human cognition and intervention.[16] This warrants a closer look, especially with regard to the emergence of such risks when nuclear weapons are involved.

[11] On the example of Harpy in particular and a comprehensive discussion of autonomy in weapon systems in general see Scharre, P., *Army of None: Autonomous Weapons and the Future of War* (W. W. Norton & Co.: New York, 2018).

[12] See the overview in Amoroso, D. et al., *Autonomy in Weapon Systems: The Military Application of Artificial Intelligence as a Litmus Test for Germany's New Foreign and Security Policy*, Heinrich-Böll-Foundation Publication Series on Democracy no. 49 (Heinrich-Böll-Stiftung: Berlin, May 2018).

[13] Amoroso et al. (note 12), pp. 23–31.

[14] On this notion see Sparrow, R., 'Killer robots', *Journal of Applied Philosophy*, vol. 24, no. 1 (Feb. 2007), pp. 62–77; Asaro, P., 'On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making', *International Review of the Red Cross*, vol. 94, no. 886 (summer 2012), pp. 687–709; and International Committee of the Red Cross (ICRC), *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?* (ICRC: Geneva, Apr. 2018).

[15] Scharre, P., *Autonomous Weapons and Operational Risk* (Center for New American Security: Washington, DC, Feb. 2016).

[16] Altmann, J. and Sauer, F., 'Autonomous weapon systems and strategic stability', *Survival*, vol. 59, no. 5 (Nov. 2017), pp. 117–42.

III. AI, machine learning and nuclear risk

Prospects of AI and machine learning in nuclear weapon systems

In the highly sensitive and notoriously conservative nuclear sector the limits to the possible automation of processes and autonomy in weapon systems are even clearer than in the conventional realm. After all, unlike in the case of conventional weapons, in the nuclear realm the final use decision invites nothing less than existential risk. Thus, it is relatively safe to assume that in the nuclear realm critical functions and final decisions will not be fully automated (at least for the foreseeable future, hopefully). Having said that, it is worth mentioning that automating nuclear decision-making is not totally unthinkable, the prime historic example being the Soviet Dead Hand system, the status and full scope of which is still unknown.[17]

The ultimate cautionary tale related to the automation of nuclear systems is provided by the example of Lieutenant Colonel Stanislav Petrov, who in 1983 called into question an alert of the Soviet early-warning system announcing a US nuclear attack.[18] It is widely acknowledged that Petrov's decision not to report this alert up his chain of command prevented a probable nuclear escalation. Petrov later explained his decision—which turned out to be correct because the alert was false—by stating that the Soviet warning system was new, that the small number of US missiles it reported did not make sense for a first strike and that his gut feeling made him doubt the authenticity of the alarm. Human judgment, as the example of Petrov shows, includes the ability to evaluate and combine numerous subtle contextual sources of information. As stated above, current AI and machine learning systems are only capable of coping with narrow and clearly defined tasks. Human-level decision-making competence as displayed by Petrov will not be reproducible in machines in the foreseeable future.

Nuclear risk and AI and machine learning in conventional weapon systems

Even though the proverbial push of the nuclear button will not be delegated to a machine anytime soon, the rush to introduce AI and machine learning in military applications risks increasing instability, including nuclear instability. One reason for this is the familiar problem of entanglement between the conventional and the nuclear realms, in particular non-nuclear threats to nuclear weapons and their command, control, communications and intelligence (C3I) systems.[19] As the capabilities of conventional arms increase, it becomes more feasible to use them to hold nuclear assets at risk. Autonomy in conventional weapon systems is one such

[17] Hoffman, D. E., *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy* (Anchor Books: New York, 2009). On the Dead Hand system see also chapters 5 and 8 in this volume.

[18] Rosenbaum, R., *How the End Begins: The Road to a Nuclear World War III* (Simon and Schuster: London, 2011), p. 7; and Blair, B. G., *The Logic of Accidental Nuclear War* (Brookings Institution: Washington, DC, 1993), p. 181.

[19] Acton, J. M. (ed.), *Entanglement: Chinese and Russian Perspectives on Non-nuclear Weapons and Nuclear Risks* (Carnegie Endowment for International Peace: Washington, DC, 2017).

advanced capability, thus feeding into the increasing entanglement and, in turn, increasing strategic instability.

One concrete example would be the deployment of stealthy UAVs and the use of swarming. Perdix is a swarming test programme currently pursued by the US Air Force. In the future, UAV swarms might facilitate the search for dispersed mobile missile launchers.[20] Another example is the use of maritime autonomous systems to hunt for nuclear-powered ballistic missile submarines (SSBNs) armed with nuclear weapons. A programme funded by the US Defense Advanced Research Projects Agency (DARPA) has resulted in the development of an autonomous trimaran, *Sea Hunter*, which is currently being tested by the US Navy.[21] Its ability to detect and pursue SSBNs could potentially limit the second-strike capabilities of other nuclear powers.

To be sure, these capabilities are still just emerging and neither Perdix nor *Sea Hunter*, nor their successors, will single-handedly destabilize the global nuclear order. Also, the hypothesis that systems such as *Sea Hunter* would render the oceans 'transparent', virtually nullifying the utility of SSBNs as a reliable second-strike asset, is hotly debated.[22] Nevertheless, just the perception that nuclear capabilities face new risks is bound to sow distrust between nuclear-armed adversaries. Moreover, a system such as *Sea Hunter* demonstrates how autonomous weapon technologies are expediting the completion of the targeting cycle, thus putting the adversary under additional pressure and potentially provoking 'use it or lose it' scenarios with regard to a nuclear second-strike capability.[23]

The entanglement problem is further aggravated by an increasing political willingness to use nuclear means to retaliate against non-nuclear attacks on early-warning and control systems or the nuclear weapons themselves. The USA signalled in its 2018 Nuclear Posture Review that it may, in future, respond with nuclear means to significant, non-nuclear strategic attacks (moving away from 'single purpose' nuclear deterrence).[24] Russia has already held this position for some time due to the USA's advantage in conventional weapons technology. Now that it is mirrored by the USA, there are likely to be further adverse effects on the stability of relations between the two largest nuclear powers.[25]

Nuclear risk and AI and machine learning in information operations

In addition to exacerbating the entanglement problem, some AI and machine learning techniques are evoking other—also long-standing but less immediate—

[20] Kallenborn, Z. and Bleek, P. C., 'Swarming destruction: drone swarms and chemical, biological, radiological, and nuclear weapons', *Nonproliferation Review*, published online 2 Jan. 2019.

[21] Trevithick, J., 'Navy's Sea Hunter drone ship has sailed autonomously to Hawaii and back amid talk of new roles', The Drive, 4 Feb. 2019.

[22] Brixey-Williams, S., 'Will the Atlantic become transparent?', 2nd edn, British Pugwash, Nov. 2016.

[23] Altmann and Sauer (note 16), pp. 130–31.

[24] US Department of Defense (DOD), *Nuclear Posture Review* (DOD: Washington, DC, Feb. 2018), p. 21.

[25] Trenin, D., 'Russian views of US nuclear modernization', *Bulletin of the Atomic Scientists*, vol. 75, no. 1 (2019), pp. 14–18.

risks to strategic stability, especially with regard to miscalculation and misperception. This includes the manipulation of the information landscape in which political and military decisions about nuclear weapons take place.

A prime example here is the deep-learning technique used to produce exceptionally accurate but false still and moving images, so-called deep fakes, especially deep fake videos generated in real time and distributed online for manipulative purposes.[26] Unfortunately, in the current era the US president's primary communications channel is the Twitter social media service, so-called fake news is running rampant on the Internet and North Korea is a nuclear-armed state. Consequently, deep fakes have added a new twist to the existing risk of manipulation, misperception and possible unintended escalation. It goes without saying that the capability to generate deep fakes is well within the range of various non-state actors.[27]

IV. Conclusions

The entry of AI and machine learning into the nuclear age comes with a reminder that nuclear risk reduction is more pertinent than ever. After all, most of the risks exacerbated by these new technologies are old and well known. But so too are some of the possible solutions to mitigate these risks. No-first-use doctrines and a lowering of the alert status of nuclear arsenals, for example, would buy valuable time during a crisis and allow for a closer evaluation of the signals received, and so prevent escalation due to miscalculation and misperception.

To paraphrase Max Tegmark, the nuclear age is a race between humankind's potential to destroy itself and its capability to avert that catastrophe.[28] Racing blindly down the path toward 'smarter' weapons, with nuclear risks remaining as inadequately addressed as they are now, might well turn the military applications of AI and machine learning into a shortcut to Armageddon.

[26] Schellmann, H., 'Deepfake videos are getting real and that's a problem', *Wall Street Journal*, 15 Oct. 2018.

[27] Brundage, M. et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation* (Future of Humanity Institute et al.: Oxford, Feb. 2018), p. 49.

[28] Tegmark, M., 'The wisdom race is heating up', Future of Life Institute, 7 Jan. 2016.

11. The destabilizing prospects of artificial intelligence for nuclear strategy, deterrence and stability

JEAN-MARC RICKLI*

The impact of artificial intelligence (AI) on nuclear weapons, strategy and deterrence is becoming of growing interest and growing concern for many observers of international security. AI is considered by many to be the new silver bullet of future warfare and a technology that will profoundly transform global power. This was clearly acknowledged by Russian President Vladimir Putin, who noted that 'Whoever becomes leader in this sphere [AI] will become the ruler of the world'.[1]

This essay briefly highlights the potential impact that AI could have on strategic stability (in section I) and nuclear deterrence (in section II). For the sake of clarity, strategic stability refers to the absence of incentives to use nuclear weapons first (in pre-emptive attacks) and the absence of incentives to build up those forces. In addition to a credible deterrent that is based on the ability to retaliate after an enemy attack, nuclear stability also requires assurance and reassurance.[2]

I. How AI could threaten nuclear stability

Impact on human decision-making

One of the key characteristics of algorithm-based systems is that they work at superhuman speeds when processing data and executing a specific task (i.e. milliseconds). Several recent studies have demonstrated that, although algorithm-based systems are still limited in what they can learn, once such as system has learned how to accomplish a specific task, it does it better and faster than a human being.[3] Project Maven, in which the US multinational technology company Google sought to automate image recognition of real-time footage from unmanned aerial vehicles (UAVs) for the United States Department of Defense (DOD), is an early manifestation of how AI could be used in military decision-making by

[1] '"Whoever leads in AI will rule the world": Putin to Russian children on Knowledge Day', RT, 1 Sep. 2017; and Maggio, E., 'Putin believes that whatever country has the best AI will be the rule of the world', Business Insider, 4 Sep. 2017.

[2] Lohn, A. J. and Geist, E., 'Will artificial intelligence undermine strategic stability?', *Bulletin of the Atomic Scientists*, 30 Apr. 2018.

[3] E.g. Wood, J., 'This AI outperformed 20 corporate lawyers at legal work', World Economic Forum, 15 Nov. 2018.

* The author would like to thank Bérangère Barthelmé and Alexander Jahns for conducting background research.

relieving the burden of data processing from human operators.[4] The objective of the programme was to fight a violent non-state actor by relying on algorithms to identify 'the weapons and tools' of an insurgency and thus allow soldiers to process data two or three times faster.[5] Project Maven's software is built on top of the open-source library TensorFlow, which makes it very difficult to build in proprietary constraints on the code. This implies that 'once the [DOD] has a trainable algorithm on hand, it can continue to develop and refine its object-recognition AI as it chooses'.[6]

The growing influence of automation in military decision-making raises several concerns, especially when it comes to nuclear deterrence and, hence, nuclear strategic stability. Indeed, human decisions based on data collected and analysed by a machine may be influenced in ways that the operator is unaware of. This has to do with several factors. The current nature of algorithm-based systems is comparable to a black box where it is impossible to retrace the decision-making process of the algorithms.[7] This lack of traceability in decision-making can create a real dilemma as the operator has to trust the outcome presented by the machine. Yet algorithms reproduce the biases of the data they rely on to learn their tasks. It is therefore impossible to exclude a risk of inadvertent escalation or at least of instability if the algorithm misinterprets and misrepresents the reality of the situation.

These examples show that AI does not need to be weaponized to represent a challenge for nuclear deterrence. Algorithm-based systems in military decision-support functions such as 'AI advisers'—which can assess a nuclear threat and plan the best response in the short time available—will have an enormous impact on conflict escalation management.[8] Combined with current developments in missile technologies and hypersonic missiles, they could create an ecosystem that could drastically shorten the already short decision-making time to respond to a nuclear attack—which is currently estimated to be around 30 minutes.[9]

Unpredictability and vulnerabilities of AI technology

Although research programmes such as the Explainable AI programme of the US Defense Advanced Research Projects Agency (DARPA) are under way, the black box nature of algorithm-based systems represents a huge accountability

[4] The project is reportedly due to end in 2019 following protests by Google employees. Conger, K., 'Google plans not to renew its contract for Project Maven, a controversial Pentagon drone AI imaging program', Gizmodo, 1 June 2018. On Project Maven see also chapters 5, 6 and 10 in this volume.

[5] Atherthon, K. D., 'Targeting the future of the DoD's controversial Project Maven initiative', C4ISRNET, 27 July 2018.

[6] Atherthon (note 5).

[7] Kuang, C., 'Can AI be taught to explain itself?', *New York Times*, 21 Nov. 2017.

[8] Hornigold, T., 'How will artificial intelligence affect the risk of nuclear war?', Singularity Hub, 28 May 2018.

[9] Macias, A., 'Here's what the US should do if Russia launches a nuclear attack, according to the top American nuclear commander', CNBC, 21 Mar. 2018.

challenge for the military.[10] On the surface, the uncertainty regarding a decision made by such a system, because of the near-impossibility of understanding the different steps leading to it, could be seen by some militaries as an advantage, providing some plausible deniability. However, this also makes the weapon less predictable—a characteristic that traditional military organizations abhor. Predictability is indeed a cornerstone of traditional military organizational culture and so it is likely that weapons equipped with machine learning algorithms will first be adopted by non-traditional actors or organizations whose organizational culture is much more responsive to disruption.[11]

The accountability challenge will grow exponentially if machine learning algorithms are used in lethal autonomous weapon systems (LAWS) and it will grow even more if, in a theoretical case for now, machine learning algorithms are ever fitted into nuclear weapons. In these scenarios, similar to what happens in financial flash crashes, failure modes are imaginable that result in unexpected situations or behaviours.[12] An early illustration of such a scenario—although not in the nuclear domain—happened during the 2016 DARPA Cyber Grand Challenge, the first hacking contest in which autonomous 'capture the flag' systems faced each other. During this competition, one autonomous system gave up in the middle of the contest, while another repaired some damage but, in doing so, crippled the machine that it was meant to protect.[13]

The risk of a nuclear war due to a failure of nuclear weapon control systems caused by an algorithm-based system is an unlikely scenario as it would imply that the decision to launch a nuclear weapon is fully automated, whereas states want humans to retain control of this decision. However, an accidental escalation resulting from incorrect information provided by an algorithm is a far more likely scenario that will have to be taken into account if nuclear systems are ever equipped with machine learning algorithms.

This risk will increase even more if adversaries are able to provide forged data or to manipulate an algorithm through a black box attack. The latter refers to a situation where different techniques (e.g. training a substitute model or using generative adversarial networks, which pit neural networks against each other) are used to work out the machine learning models of another algorithm-based system.[14] In this situation, it would be possible to manipulate the data of an adversary and therefore fool its defence system. The possibility of fooling algorithms through adversarial attacks that will cause an image to be miscategorized, for instance, or

[10] Gunning, D., 'Explainable Artificial Intelligence', Defence Advanced Research Projects Agency (DARPA), [n.d.].

[11] Horowitz, M. C., 'Artificial intelligence, international competition, and the balance of power', *Texas National Security Review*, vol. 1, no. 3 (May 2018), pp. 37–57.

[12] Akioyamen, P., 'Neural networks and deep learning—the revival of HFT', Medium, 22 July 2018; and Turchin, A. and Denkenberger, D., 'Classification of global catastrophic risks connected with artificial intelligence', *AI & Society*, published online 3 May 2018.

[13] Metz, C., 'Hackers don't have to be human anymore. This bot battle proves it', *Wired*, 8 May 2016.

[14] Botta, A., 'Getting to know a black-box model', Towards Data Science, 24 July 2018; Giles, M., 'The GANfather: the man who's given machines the gift of imagination', *MIT Technology Review*, 21 Feb. 2018; and Hu, W. and Tan, Y., 'Generating adversarial malware examples for black-box attacks based on GAN', arXiv, 1702.05983, 20 Feb 2017.

the wrong synthetic data to be provided is real and is indeed a growing concern for data security specialists.[15]

II. How AI could have an impact on nuclear deterrence

AI to undermine the second-strike capability of nuclear states

In the short term, the major destabilizing impact of AI on nuclear deterrence is in the combination of autonomy and the fusion of all kind of sensors that will make or appear to make the survival of second-strike capabilities less likely and hence reduce strategic stability. Mutually assured destruction (MAD) relies on the assumption that the potential attacker has no incentive to launch a nuclear strike if the defender can guarantee a retaliatory strike.

It was not until the development of nuclear-powered ballistic missile submarines (SSBNs) that nuclear deterrence became stable.[16] Indeed, before submarines were equipped with nuclear missiles, a theoretical situation could be imagined where a potential adversary would launch a pre-emptive attack against all nuclear bombers and nuclear silos of its opponent and thus annihilate the opponent's retaliatory nuclear capabilities before launching its own nuclear attack. With the introduction of submarines equipped with nuclear missiles, deterrence became stable as it was impossible to wipe out all of the opponent's SSBNs as the location of each is known only to the commander of the boat.

This might dramatically change with progress in capacities for AI-enabled tracking and targeting of adversaries' nuclear weapons.[17] The sacrosanct assumption that SSBNs are immune to a pre-emptive strike could disappear due to the contributions of AI to intelligence, surveillance and reconnaissance (ISR) systems and the ability of offensive unmanned underwater vehicles (UUVs) to chase SSBNs.[18] However, the technology is not yet mature. It is worth mentioning that this assumption had already been challenged by the rise of cyber vulnerabilities in missile launchers. Indeed, a 2017 study demonstrated that the British Trident system of SSBNs was not immune to hacking.[19]

The ability of AI to make predictions based on the fusion of disparate sources of information that enable it to find and target missiles stored in silos and on

[15] Goodfellow, I., 'Attacking machine learning with adversarial examples', OpenAI, 24 Feb. 2017; Oberhaus, D., 'Researcher created fake "master" fingerprints to unlock smartphone', Motherboard, 15 Nov. 2018; and Brundage, M. et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation* (Future of Humanity Institute et al.: Oxford, Feb. 2018). On adversarial attacks see also chapter 13 in this volume.

[16] Stanhope, M., 'Lessons on strategic stability and SSBNs from the cold war', The Interpreter, 12 Dec. 2014.

[17] Boulanin, V., 'AI and nuclear weapons—promise and perils for nuclear stability', AI & Global Governance, United Nations University, Centre for Policy Research, 7 Dec. 2018.

[18] Borchert, H., Mahon, D. and Kraemer, T., 'Leveraging undersea autonomy for NATO: allies must work together to avoid fraction', *Cutting the Bow Wave*, 2016, pp. 50–53; and Snyder, R., 'The future of the ICBM force: should the least valuable leg of the triad be replaced?', Policy White Paper, Arms Control Association, Mar. 2018, p. 2.

[19] Abaimov, S. and Ingram, P., *Hacking UK Trident: A Growing Threat* (British American Security Information Council (BASIC): London, June 2017).

aircraft, submarines or trucks is growing. This 'could enable the development of strategically destabilizing threats to the survivability' of missile launchers and especially of mobile intercontinental ballistic missile (ICBM) launchers—the cornerstone of nuclear deterrence.[20] With such capabilities, the threat of retaliation could be ruled out, and thus invite a first strike—a very destabilizing prospect indeed.

Perceived versus actual capabilities

By its very nature, nuclear deterrence is highly psychological and relies on the perception of the adversary's capabilities and intentions. Here, however, lies the trickiest and most deceitful destabilizing influence that current advances in AI have on nuclear deterrence: the simple misperception of the adversary's AI capabilities is destabilizing in itself. As Edward Geist and Andrew Lohn of the Rand Corporation rightly observe, 'the effect of AI on nuclear strategy depends as much or more on adversaries' perceptions of its capabilities as on what it can actually do'.[21]

The current potential for misperceptions is already important. For instance, in order to prevent the US nuclear fleet from being attacked asymmetrically, the US DOD Defense Science Board argued for the USA to 'be more proactive and complement [its] submarine force with other capabilities', such as autonomous UUVs and sensor networks.[22] In March 2018 President Putin, in his annual address to the Russian Federal Assembly, stated that

> Now, we all know that the design and development of unmanned weapon systems is another common trend in the world. As concerns Russia, we have developed unmanned submersible vehicles that can move at great depths (I would say extreme depths) intercontinentally, at a speed multiple times higher than the speed of submarines, cutting-edge torpedoes and all kinds of surface vessels, including some of the fastest. It is really fantastic. They are quiet, highly manoeuvrable and have hardly any vulnerabilities for the enemy to exploit. There is simply nothing in the world capable of withstanding them.
>
> Unmanned underwater vehicles can carry either conventional or nuclear warheads, which enables them to engage various targets, including aircraft groups, coastal fortifications and infrastructure.[23]

Such rhetoric directly fuels potential misperceptions about capabilities and intentions.

With any new weapon system, there is a lot of speculation about what the system can do. For instance, the first sea trials of the new nuclear-capable underwater vehicle Poseidon—the system announced by Putin in March 2018—allegedly started in July 2018, but the outcomes of these tests remain open for

[20] Groll, E., 'How AI could destabilize nuclear deterrence', *Foreign Policy*, 24 Apr. 2018.

[21] Geist, E. and Lohn, A. J., *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (Rand Corporation: Santa Monica, CA, 2018), p. 1.

[22] US Department of Defense (DOD), Defense Science Board, *Report of the Defense Science Board Summer Study on Autonomy* (DOD: Washington, DC, June 2016), p. 61.

[23] President of Russia, 'Presidential address to the Federal Assembly', 1 Mar. 2018. See also 'Status-6/Kanyon–Ocean Multipurpose System', GlobalSecurity.org, [n.d.].

speculation.[24] With the current hype surrounding the achievements of AI, the simple perception that it is a successful technology available to an adversary can in itself be destabilizing. According to one commentator, in nuclear deterrence 'misconceptions about what artificial intelligence can do can be just as dangerous as AI itself'.[25] This represents the biggest current challenge of AI to the stability of nuclear deterrence as nuclear powers do not need to actually acquire autonomous capabilities in order to challenge strategic stability. Thus, the nuclear powers should consider, with the highest priority, communicating clearly and accurately about their AI capabilities in order to avoid a global AI arms race that has the potential to completely upset the current nuclear strategic balance.

An AI arms race

A group of advanced military powers comprising Australia, Israel, the Republic of Korea (South Korea), Russia and the USA among others has consistently blocked any progress towards a new international treaty or political declaration to ban fully autonomous weapon systems during the various meetings held since 2014 in the framework of the 1980 Convention on Certain Conventional Weapons (CCW Convention).[26] Given this, there should be no remaining illusions about the looming arms race in AI technologies that will characterize the future global power relationship. For instance, in a study on autonomy, the US Defense Science Board concluded that the DOD 'must accelerate its exploitation of autonomy—both to realize the potential military value and to remain ahead of adversaries who also will exploit its operational benefits'.[27] In July 2017 China set the goal of becoming the leader in the field of AI by 2030, to challenge US dominance.[28] The words of President Putin quoted at the start of this essay also made clear the position of Russia. Thus, the development of weapons relying on AI and autonomy will be a key characteristic of arms races in the 21st century. For now, the development of autonomous nuclear weapons is not planned by any nuclear power; but, as demonstrated above, AI in nuclear command-and-control or surveillance systems cannot be discounted.

However, unlike the nuclear arms race, this AI arms race will probably also involve many more actors, and these will not be restricted to states.[29] Due to

[24] Gady, F.-S., 'Russia begins sea trials of nuclear-capable Poseidon underwater drone', *The Diplomat*, 21 July 2018. See also Insinna, V., 'Russia's nuclear underwater drone is real and in the Nuclear Posture Review', *Defense News*, 12 Jan. 2018.

[25] Hornigold (note 8).

[26] Delcker, J., 'How killer robots overran the UN', *Politico*, 12 Feb. 2019; and Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects (CCW Convention), opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983.

[27] US Department of Defense (note 22), p. 1.

[28] Cadell, C. and Jourdan, A., 'China aims to become world leader in AI, challenges U.S. dominance', Reuters, 20 July 2017; and Chinese State Council, 'China issues guideline on artificial intelligence development', 20 July 2017.

[29] Rickli, J.-M., 'The impact of autonomy and artificial intelligence on strategic stability', *UN Special*, no. 781 (July–Aug. 2018), pp. 32–33.

the scalability, efficiency and ease of diffusion of AI systems, the cost (in terms of resources, manpower and psychological distance) of carrying out attacks will be lower, potentially increasing the number of malevolent actors as well as the number of attacks that can be carried out.[30]

UAV technology exhibits similar proliferation features to those technologies that rely on autonomy. Although the first weaponized UAVs appeared during the 1959–75 Viet Nam War, it was not until the end of the 2000s that this technology became available to non-state actors. Non-state actors such as Hezbollah, the Islamic State group and the Houthi movement have now also acquired these capabilities off-the-shelf.[31] For instance, during the 2017 Battle of Mosul, Islamic State mounted 40-millimetre grenades onto UAVs to drop them on Iraqi Government positions, killing up to 30 Iraqi soldiers in a single week.[32]

Emerging technologies tend to exhibit similar characteristics. They are expensive to first develop, but then the price drops dramatically once they are commercialized. For algorithm-based systems, this could go even faster as most are developed using open source software; even when they are not, once an algorithm has been developed the price of reproducing it is almost non-existent. This will call for new ways to approach arms control with mechanisms that both span the divide between conventional and nuclear munitions and address horizontal proliferation to lesser powers and vertical proliferation to non-state actors.[33] This will also change the character of warfare since, as access to weapon technologies is democratized, states and non-state actors will be tempted to use surrogates to fight on their behalf.[34]

III. Conclusions

The prospects of the application of AI in nuclear deterrence and nuclear strategy has the potential to reduce strategic stability. New AI technologies will introduce new offensive threats, as AI systems can complete tasks more successfully or take advantage of vulnerabilities in other AI systems, including autonomous weapon systems. Threats might be altered by AI technologies, making attacks typically more effective, more finely targeted, more difficult to attribute and more likely to exploit vulnerabilities in the AI systems of the adversary.

Applications of AI in the decision-support systems that deal with the use of nuclear weapons and in the tracking and targeting of an adversary's launchers will dramatically improve targeting accuracy and, probably, increase the tempo of operations. These too will disrupt nuclear stability. Of even more concern, however, is the fact that just the perception of these capacities by an adversary

[30] Brundage et al. (note 15).

[31] Rickli, J.-M., *The Economic, Security and Military Implications of Artificial Intelligence for the Arab Gulf States* (Emirates Diplomatic Academy: Abu Dhabi, Nov. 2018).

[32] Chovil, P., 'Air superiority under 2000 feet: lessons from waging drone warfare against ISIL', War on the Rocks, 11 May 2018.

[33] Klare, M. T., 'The challenges of emerging technologies', *Arms Control Today*, vol. 48, no. 10 (Dec. 2018).

[34] Krieg, A. and Rickli, J.-M., *Surrogate Warfare: The Transformation of War in the Twenty-First Century* (Georgetown University Press: Washington, DC, 2019).

will be destabilizing in itself. To make matters worse, the ease of proliferation of AI technologies to both states and non-state actors will add an additional layer of complexity and probably force a rethink of traditional concepts of nuclear deterrence.[35]

[35] Chertoff, P., *Perils of Lethal Autonomous Weapons Systems Proliferation: Preventing Non-State Acquisition*, Strategic Security Analysis no. 2 (Geneva Centre for Security Policy: Geneva, Oct. 2018).

12. The impact of unmanned combat aerial vehicles on strategic stability

JUSTIN BRONK

Combat aircraft are usually the tool of choice for governments that wish to reassure, deter and signal with military force. As such, for decades the judgement of human pilots has been an important factor in strategic stability, from cold war alert-patrol clashes and overflight interceptions to encounters on tense national borders such as that between India and Pakistan and in the Taiwan Strait. However, with advances in automation and powerful operational requirements, there is now pressure on first-tier air forces around the world to develop and deploy unmanned combat aerial vehicles (UCAVs). These are pilotless combat aircraft designed with a high degree of survivability and lethality for use in contested airspace.

This essay discusses the factors that are pushing nuclear-armed states and other major military powers towards the development and acquisition of UCAVs. It starts with an assessment of the comparative advantage of UCAVs (section I), covering both the benefits of UCAVs themselves and the weaknesses of existing unmanned aerial vehicles (UAVs). The essay then reviews the current state of UCAV technology and the extent to which it has been adopted (section II). Finally, it concludes by looking at the requirement for autonomy in UCAVs and the need for a discussion on its responsible use (section III).

I. The comparative advantage of UCAVs

The unsuitability of existing unmanned aerial vehicles for high-intensity warfare

The predominant type of UAV used by air forces around the world today is the medium-altitude long-endurance (MALE) remotely piloted aircraft. The US MQ-9 Reaper and the Chinese Wing Loong series are in many ways synonymous with the use of UAVs (or 'drones') in modern warfare. However, these systems are not suitable for use in high-intensity conflict due to their lack of self-defence capabilities and their reliance on real-time remote control via satellite communications (satcom) links by aircrew sitting in ground stations. Satcom is comparatively simple to interrupt, deny, intercept or spoof due to the distances at which it must operate compared to local jamming sources. This was demonstrated in 2011 when Iran captured a stealthy and highly advanced RQ-170 Sentinel UAV by overriding the command link.[1] There are also examples of insurgent groups such as Hezbollah tapping into Israeli UAV feeds.[2]

[1] 'Iran "building copy of captured US drone" RQ-170 Sentinel', BBC, 22 Apr. 2012.

[2] E.g. Grant, G., 'Hezbollah claims it hacked Israeli drone video feeds', Military.com, 10 Aug. 2010.

Satcom is likely to be extremely unreliable in the contested electromagnetic environment of a state-on-state conflict, especially when it involves advanced military powers such as China, Russia and the United States. Furthermore, there are significant limitations to the manoeuvres that a remotely piloted aircraft can perform without temporarily losing contact with its controller, which limits its defensive options once engaged by hostile air- or surface-based threats.

The operational benefits of UCAVs

UCAVs are a different concept since they would fly missions according to preprogrammed or dynamically tasked instructions, rather than being remotely flown in real time. As such, there is no requirement to train a cadre of aircrew or remote crews in the style of traditional combat aircraft or UAVs. As an example, while the British Royal Air Force has a fleet of approximately 145 Typhoon combat aircraft, only around 55 of these are generally available for frontline use.[3] This is because some are needed to train new pilots, and multiple squadrons are needed for each one at combat-deployable readiness.

A typical US combat aircraft pilot might spend 6 months out of every 24 deployed on operations or, during periods of force strain, up to 6 in every 18 or even 12 months.[4] The rest of the time is spent either resting and recuperating after combat tours or training and retraining to maintain skills and expand the qualifications for younger pilots in each unit. This not only means that two to three squadrons are needed for each one currently deployed or ready to deploy, but that the available fatigue life (i.e. the flying hours for which a given aircraft is certified to be used in its lifecycle) of manned combat aircraft is largely used up in training and currency-maintenance sorties rather than on operations.

With UCAVs, none of the extra squadrons are theoretically required, and the majority of sorties actually flown can be on operations since pilot training and currency (i.e. the legal requirement to stay 'current') are not an issue. This would greatly expand the combat power represented by each aircraft purchased compared to a combat aircraft of the same survivability and lethality. Furthermore, commonality for pilots (i.e. being trained and currently qualified on one aircraft in a fleet allowing a pilot to fly all the others) is not required across a UCAV fleet. Thus, such a force could offer far more design flexibility while in service than a piloted aircraft fleet in response to changing threat outlooks.

Production of UCAVs for force expansion or to replace losses in combat would be a matter of industrial capacity and supply chain management. While this would be difficult and expensive, it would not be impossible at reasonably short notice. Since training combat aircraft aircrew takes years and requires a supply of experienced instructor pilots, no such rapid force expansion or replenishment of manned aircraft is likely to be possible at short notice or following serious losses.

[3] British Ministry of Defence, Air Command Secretariat, Freedom of Information Request no. 2017/1418, 24 Feb. 2017; and British Government, *Securing Britain in an Age of Uncertainty: The Strategic Defence and Security Review* (Stationery Office: London, Oct. 2010).

[4] Losey, S., 'Air Force deployment tempo brings new kinds of strains', *Air Force Times*, 29 Mar. 2016.

II. The state of UCAV technology and its adoption

Technological requirements: UCAVs and AI technology

UCAVs capable of performing key missions would not require general AI or any particularly advanced automation technologies beyond the levels that have already been proven in combat air and other sectors.[5] These key missions would include defensive counter air (DCA), offensive counter air (OCA), suppression or destruction of enemy air defences (SEAD/DEAD), and deep strike against critical targets in a high-intensity conflict scenario.

The detection and destruction of active hostile surface-to-air missile (SAM) radars or combat aircraft in a major war is not something that requires subtle judgements in terms of international humanitarian law and estimates of proportionality or collateral damage. In particular, SEAD/DEAD missions would be conducted with a heavy reliance on munitions that already possess significant autonomy in target discrimination since they must be launched from outside the direct sensor ranges of launch aircraft.[6]

The UCAV programmes of nuclear-armed states

Together, the greater efficiencies, greater force densities for a given fleet size and greater speed of reaction in flight to threats compared to manned aircraft mean that a UCAV force could greatly increase a state's options for both defensive and offensive use of air power against hostile states with advanced military capabilities. Given the extent to which both China and Russia—not to mention smaller nuclear-armed states such as the Democratic People's Republic of Korea (DPRK, or North Korea) and Pakistan—rely on potent ground-based air defences to protect their critical national assets, a swing towards an offensive advantage in air power would have significant implications for strategic stability, potentially increasing pressures to 'use or lose' nuclear capabilities due to fears of a disarming first strike.

The USA's X-45 programme proved the capability for UCAV prototypes to dynamically detect, prosecute and attack SAM threats as a multi-aircraft formation as early as 2005.[7] In this scenario the X-45 UCAVs were operating according to their preprogrammed rules of engagements and situational awareness created by pooling their respective sensor data. With more than a decade of progress in the critical fields of data processing, automation and sensor fusion since the X-45 programme, it would seem safe to assume that current combat aircraft technology could already produce combat-capable UCAVs with a capacity for cooperative automated warfare. Without human endurance concerns

[5] Bronk, J., *Next Generation Combat Aircraft: Threat Outlook and Potential Solutions* (Royal United Services Institute: London, Nov. 2018).

[6] Bronk, J., 'Eastern Europe—a no-fly zone for the West?', *RUSI Defence Systems*, vol. 18, 13 May 2016.

[7] On the X-45 and X-47B programmes and UCAV technology more generally see Rogoway, T., 'The alarming case of the USAF's mysteriously missing unmanned combat air vehicles', The Warzone, The Drive, 9 June 2016.

and the need to carry heavy, complex and radar signature-increasing cockpits and clear canopies, UCAVs also offer longer range and greater persistence 'on station' (i.e. within the designated mission area) compared to manned combat aircraft of a similar size. They can also be used in high-risk scenarios without the need for combat search-and-rescue support in case an aircrew is downed in hostile territory—not to mention the lower risk to life without human occupants. While the US Air Force has stated that it sees no place for unmanned aircraft with a nuclear mission, its new long-range bomber, the B-21 Raider, is being developed with the capacity to operate without crew in future conventional missions, making it a potentially highly automated dual-capable combat aircraft.[8]

China is certainly pursuing UCAV capabilities with its Dark Sword, Sharp Sword and CH-7 stealth UCAV prototypes, which were glimpsed through the carefully managed 'leak' of photos and even an appearance by the CH-7 demonstrator model at the Zhuhai Airshow in 2018.[9] The USA has also proved that it can design and test prototype UCAVs, with the X-45 and X-47 programmes, as has the United Kingdom with Taranis and France with nEUROn. Technologically speaking, the genie is essentially out of the bottle, although India and Russia continue to have problems with the development of stealthy and highly automated combat aircraft due to industrial limitations. What remains to be seen is not whether UCAVs are developed but whether these aircraft are produced and introduced into service, at scale and in view of the public.

III. The need for UCAVs to be autonomous

Autonomy: an operational requirement

For UCAVs to make sense as an investment decision for advanced air forces, they must be capable of detecting, classifying, prioritizing and engaging targets with lethal force according to preset mission tasks and rules of engagement without real-time human control. This is because, in a high-intensity scenario in a highly contested area, it is likely that the data or satcom link to a UCAV would be at least intermittently jammed or disrupted. For any UCAV design that is intended to have utility in a conflict involving nuclear weapons, the likelihood of anti-satellite (ASAT) warfare in the opening stages of a nuclear exchange would make the ability to operate at long-range without satcom even more critical. In other words, UCAVs must be lethal autonomous weapon systems (LAWS) almost by logical necessity.[10] However, this does not preclude them from being developed with the latent capability to operate autonomously while remaining subject to a human 'yes or no' weapon-release authority in all scenarios where connectivity is possible. There are already many stand-off munitions (e.g. cruise missiles) that

[8] See e.g. Sayler, K. and Scharre, P., 'The B-21 bomber should be unmanned on Day 1', Defense One, 31 May 2016.

[9] Kang, D. and Bodeen, C., 'China unveils stealth combat drone in development', Associated Press, 7 Nov. 2018.

[10] Bronk (note 5).

possess comparable automatic target-detection, classification, prioritization and attack capabilities when in 'war mode' but which are not current employed in that way (e.g. most long-range anti-ship missiles and anti-radiation missiles).

The need for a debate on responsible use of autonomy in UCAVs

All this does not mean that UCAVs should replace piloted combat aircraft, since the human capacity to understand complex and nuanced situations in combat scenarios short of high-intensity warfighting remains essential. In a nuclear mission—where the fate of billions potentially rests on split-second decisions, and where complex contextual understanding has historically brought humanity back from the brink on more than one occasion—there are strong arguments for maintaining piloted aerial delivery platforms.[11]

However, with the return of great power competition, the almost irresistible series of advantages for advanced warfighting will push air forces towards UCAVs as part of their force mix. Therefore, the ethical and legal debates around their development and use in democratic countries need to be held now.

On the one hand, these discussions on UCAVs should consider legal issues around defining meaningful human control of unmanned assets in communications-denied environments, requirements for certification under Article 36 of the 1977 Additional Protocol I to the Geneva Conventions, and the ethical differences between employment in low- and high-intensity conflict.[12] On the other hand, a mature, more general discussion about the nature of proportionality and discrimination in high-intensity warfare—where time is critical, information is partial, stand-off ranges often outstrip the range of launch-platform sensors, and large-scale casualties and disinformation are everyday phenomena—is something that is needed as the world returns to great power competition.

The effects of UCAVs (on one or both sides) on strategic stability also need to be understood. How, for example, does the employment of highly survivable but pilotless aircraft affect geopolitical signalling through deployment of combat aircraft and airspace probing? Is the potential shooting down of a UCAV during tense encounters rendered marginally less escalatory but more likely without aircrew deaths involved? How does a potentially more capable strike force that can tolerate more losses affect counterforce and deterrence options? These questions must be discussed and modelled early to help reduce the danger for future miscalculations.

Potential adversary powers (and most probably the USA) will not wait for West European powers to make up their mind before making lethal, highly autonomous aircraft. If European members of the North Atlantic Treaty Organization (NATO), including the UK, and their partner states are to influence the construction of norms around these systems, they must acknowledge their advantages as

[11] On the imperative for human control of nuclear weapon launch decisions see chapters 5–10 and 14 in this volume.

[12] Protocol I Additional to the 1949 Geneva Conventions, and Relating to the Protection of Victims of International Armed Conflicts, opened for signature 12 Dec. 1977, entered into force 7 Dec. 1978.

well as the legal and ethical questions around their potential uses. Most of all, to participate in the experimentation and debates that will shape UCAV use in future decades, West European states must offer something that shows that they are not simply seeking to blindly restrict the use of capabilities they themselves do not have or understand.

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

SHAHAR AVIN AND S. M. AMADAE*

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.[1] 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.[2] 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 the 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 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*

[1] 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, p. 120.

[2] Thompson, N., 'Inside the apocalyptic Soviet doomsday machine', *Wired*, 21 Sep. 2009; and '"Doomsday machine": Russia's new weapon reportedly gets nuclear warhead', Sputnik, 17 May 2018. See also the other chapters, in particular chapters 5–11 and 14, in this volume.

* 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.

a security policy'.[3] 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.[4] 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 system. 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.[5] 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 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

[3] US Department of Defense (DOD), *Department of Defense Trusted Computer Systems Evaluation Criteria*, DOD Standard 5200.28-STD (DOD: Washington, DC, 26 Dec. 1985), p. 116.

[4] Anderson, R., *Security Engineering: A Guide to Building Dependable Distributed Systems*, 2nd edn (Wiley: Indianapolis, IN, 2008), chapter 26.

[5] 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. 1278–308.

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.[6]

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).[7] 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.[8]

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).[9] 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.[10] 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 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.[11] 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,

[6] Barreno, M. et al., 'The security of machine learning', *Machine Learning*, vol. 81, no. 2 (Nov. 2010), 121–48.

[7] Szegedy, C. et al., 'Intriguing properties of neural networks', arXiv, 1312.6199, version 4, 19 Feb. 2014.

[8] Geist, E. and Lohn, A. J., *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (Rand Corporation: Santa Monica, CA, 2018), p. 10.

[9] 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). On NC3 see Anderson (note 4), chapter 13. On these issues see also chapters 6–11 and 14 in this volume.

[10] On the vulnerability of machine learning to cyberattack as an obstacle to its adoption in the military sphere see also chapters 4 and 7 in this volume.

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

used for both nuclear and conventional weapons—such as satellites, intelligence gathering and logistics.[12] 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 may 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 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 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

[12] 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.

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.[13]

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.[14] 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.[15]

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.[16] Machine learning techniques and automation expedite the search for patterns of common mistakes that could lead to an exploit. Access to compiled code, 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.[17]

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

[13] 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).

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

[15] 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.

[16] Jon B. and Rich T., 'A day in the life of an NCSC vulnerability researcher', British National Cyber Security Centre, 17 Nov. 2017.

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

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.[18]

At this point, country A can profile its targets and identify the intermediary influencers it would need to engage.[19] 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

[18] 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).

[19] 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.

person.[20] 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.[21] 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.[22]

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 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

[20] 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.

[21] 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. On GANs see also chapter 2 in this volume. On the malicious use of deepfakes see also chapter 10 in this volume.

[22] Rössler, A. et al., 'FaceForensics: a large-scale video dataset for forgery detection in human faces', arXiv, 1803.09179, 24 Mar. 2018.

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.[23] 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 intelligence, surveillance and reconnaissance (ISR) architectures must be maintained at high levels of reliability.[24]

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

[23] US Government Accountability Office (note 9), 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).

[24] Acton (note 12).

Review, which presents the threat of nuclear retaliation as a deterrent against cyberattack.[25] 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.[26] 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 not offer any recommendations.[27] 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 13.1. 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.

[25] US Department of Defense (DOD), *Nuclear Posture Review* (DOD: Washington, DC, Feb. 2018).

[26] Lindsay, J. R., 'Restrained by design: the political economy of cybersecurity', *Digital Policy, Regulation and Governance*, vol. 19, no. 6 (2017), pp. 493–514.

[27] US Government Accountability Office (note 9), preface.

Table 13.1. 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; 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; counter-intelligence.	As above.

Infrastructure	Upgrade cyber-defences; use cost–benefit analysis; risk assessment to prioritize high-value assets.	Enhance industry standards; monitoring (can use ML); testing; attribution capacity.	Assess; report; recall; upgrade.
Research and development, testing, simulation, maintenance	Procurement guidelines; budget for security; expert, vetted and valued staff; redundancy.	Oversight; accountability; counter-intelligence.	Evaluate; counter; attribute.

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.

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.[28] 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.

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

[28] Stoutland and Pitts-Kiefer (note 23), p. 8.

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 weapon 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 is 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.[29] 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.[30] 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 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-armed 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.

[29] US Government Accountability Office (note 9), 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.

[30] 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.

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.[31]

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

[31] Brundage et al. (note 13).

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.

14. Mitigating the challenges of nuclear risk while ensuring the benefits of technology

ANJA KASPERSEN AND CHRIS KING*

In examining the nexus between technological innovations such as artificial intelligence (AI) and nuclear risk, it is important to keep some caveats in mind.

The first is to avoid being either too alarmist or too speculative.

Second is to recognize that the nuclear order is now multipolar. The East versus West nuclear binary has expanded to regional nuclear rivalries and even strategic triangles. Emerging technologies will probably affect each of these relationships differently and will be dependent on a variety of factors. These include geographic proximity, arsenal size and sophistication, the maturity of the strategic relationship, and technological symmetry or asymmetry.

The third caveat is the issue of technological convergence. Unlike any previous technological revolution, innovations are overlapping as never before, further increasing uncertainty. The two enabling technologies of cyber capabilities and AI stand out as particular examples.

What is certain is that technological innovations from AI to cyber capabilities and hypersonic weapons are making their way into defence and security doctrines and platforms. As the United Nations Secretary-General, António Guterres, has said, the advent of potentially destabilizing weapons would be 'worrying even in the most benign security environment', let alone in one that is characterized by mistrust, deteriorating relations and the erosion of arms control instruments, and where 'military solutions could take precedence over dialogue and diplomacy'.[1] Rather than waiting for a demonstration of the challenges posed by technology, responsible policymakers must be actively engaged now. Understanding how developments in technology can increase nuclear risk is vital to preserving the seven-decade-long norm against the use of nuclear weapons. Yet, likewise, it is also incumbent on policymakers to be alert to the possibilities of positive technological disruption, including those that will create space for new approaches to disarmament and non-proliferation, such as enhanced safeguards and verification.

This essay examines how states can work together and with new and old partners to address governance gaps and to maximize the opportunities that technologies present to make the world safer and more secure. It starts in section I by mapping the potential impacts of technological innovation on nuclear risks and by considering how to reduce those risks. In section II it then assesses the ways in which risks posed by technological innovation can be governed. Finally, in

[1] UN Secretary-General, 'Secretary-General's remarks to Turtle Bay Security Roundtable: managing the frontiers of technology', 23 Mar. 2018.

* The views expressed in this piece are those of the authors and do not reflect the official policy or position of the United Nations.

section III it considers the ways in which machine learning and other technologies can be exploited to support nuclear compliance and verification regimes.

I. Potential impacts of technological innovation on nuclear risk

As a starting point, it is worthwhile touching briefly on some of the ways in which emerging technologies could increase nuclear risks. Measures to maintain the norm against the use of nuclear weapons are grounded in various interlinked understandings that emerged from the cold war. These include acceptance of mutual vulnerability; that newer and more capable nuclear weapons undermine, not reinforce, stability; and that risk reduction that builds confidence can improve prospects for international peace and security. Recent technological innovations have the potential to undermine these understandings.[2]

Mutual vulnerability

Mutual vulnerability is predicated on the understanding that, regardless of the strength of a nuclear-armed state's first-strike capability, its nuclear-armed opponents will be able to inflict devastating responses. Enhanced intelligence, surveillance and reconnaissance (ISR) capabilities driven by emerging technologies could undermine this concept. It has been posited, for example, that autonomous technologies could expose second-strike capabilities such as nuclear-powered ballistic missile submarines (SSBNs) or road-mobile missiles.[3] For example, long-range, autonomous unmanned aerial vehicles (UAVs) that use swarming technologies to allow for constant real-time monitoring, coupled with the ability to rapidly process large amounts of data using machine learning, could facilitate the tracking of these previously hidden nuclear forces.[4]

Any future conflict between reasonably advanced actors will probably include a cyber component as each side will attempt to destroy, disrupt or confuse enemy sensors, communication and decision-making loops. The cyber dimension of future warfare will have a considerable impact on global nuclear relations and doctrines. The introduction of concepts such as so-called left-of-launch missile defence could create worrying ambiguities about cyber pre-emption, increase perceptions of transformed and weakened deterrents, and drive states towards 'use it or lose it' mentalities.

The problems often associated with offensive cyber capabilities and vulnerabilities—such as the shelf-life of exploits, timely attribution and appropriate response—take on even greater weight with the addition of nuclear consequences.

[2] Futter, A., *Hacking the Bomb: Cyber Threats and Nuclear Weapons* (Georgetown University Press: Washington, DC, 2018).

[3] This scenario is discussed in greater detail in chapters 6, 10 and 11 in this volume.

[4] See e.g. Hambling, D., 'The inescapable net: unmanned systems in anti-submarine warfare', British–American Security Information Council (BASIC) Parliamentary Briefings on Trident Renewal no. 1, Mar. 2016. The term 'unmanned' is used here for consistency with the rest of this volume. A better, ungendered term would be 'uninhabited' or 'uncrewed'.

In this context, even the perception of vulnerability and nuclear modernization can heighten perceptions of risks.[5]

New weapon systems not only increase existing risks but also introduce new vulnerabilities

Closely linked to the acceptance of mutual vulnerability is the understanding that more capable nuclear weapons provide no real advantage in overcoming said vulnerability. The advent of new technologies that can enhance the speed, stealth, accuracy and manoeuvrability of nuclear weapons seems to have reversed this understanding with potentially dangerous consequences.[6]

For example, use of machine learning and autonomous technologies in conventional systems and platforms could result in unwarranted armed responses and loss of control, leading to unintended escalation.[7] In the future, growth in both data volume and computing capability, including machine learning, could increase the speed of warfare, leading to increasingly compressed decision cycles and growing pressure on human commanders. The human–machine decision-making interface is a key concern for the possible weaponization of AI. This concern is reflected in ongoing debates at the United Nations in Geneva, including in the group of governmental experts on emerging technologies in the area of lethal autonomous weapon systems (LAWS) under the aegis of the 1980 Convention on Certain Conventional Weapons (CCW Convention).[8]

In an increasingly dense fog of war, fear of losing quickly might create incentives for rapid responses, including nuclear responses, raising the chances of miscalculation.[9]

The quest for faster, smarter, more accurate and more versatile weapons could lead to destabilizing arms races. In a world with asymmetrical military technology, nuclear-armed states could cling more tightly to their arsenals and technologically disadvantaged states may seek to acquire nuclear weapons as a more achievable deterrent.

New types of weapon technology also create new vulnerabilities. It is possible that autonomous systems could be susceptible to hacking and spoofing, possibly

[5] See e.g. Lin, H. and Zegart A. (eds), *Bombs, Bytes and Spies: The Strategic Dimensions of Offensive Cyber Operations* (Brookings Institution: Washington, DC, 2019).

[6] See e.g. Kristensen, H. M., McKinzie, M. and Postol, T. A., 'How US nuclear force modernization is undermining strategic stability: the burst-height compensating super-fuze', *Bulletin of the Atomic Scientists*, 1 Mar. 2017.

[7] This possibility is discussed in greater detail in chapters 9 and 10 in this volume.

[8] E.g. Group of Government Experts of the Parties to the CCW Convention, Report of the 2018 session of the group of governmental experts on emerging technologies in the area of lethal autonomous weapons systems, CCW/GGE.1/2018/3, 23 Oct. 2018; and Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects (CCW Convention), opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983.

[9] Scharre, P., *Army of None: Autonomous Weapons and the Future of War* (W. W. Norton & Co.: New York, 2018), p. 305.

by third parties, and there are concerns that such systems could act in unexpected ways.[10]

The addition of autonomy, for example, to nuclear delivery vehicles—as has been mooted in the Poseidon unmanned underwater vehicle (also known as Status-6) and possibly the B-21 Raider strategic bomber—could make them more vulnerable, which would undermine their predictability, increase the prospects for miscalculation and decrease stability.[11]

The challenge of malicious non-state actors is particularly acute in this regard. Current arms control regimes assume that the main threats come from states. This has been true for most of history because the capacity to wreak destruction on a massive scale has typically required an army and, usually, a large research and development budget. However, destructive capacity is in the process of being democratized, thanks to the unprecedented dispersal of technological capabilities and skills.

Regarding predictability, even machine learning-based autonomous systems that achieve high rates of accuracy in training data to assess the best course of action can produce unexpected results and behaviours.[12] Such results in this context could include an autonomous nuclear delivery vehicle that cannot be recalled once deployed.

Risk reduction

Nuclear risk reduction during and since the cold war has been achieved through the painstaking construction of an interwoven safety net of political initiatives, transparency and confidence-building measures, and legally binding treaties and instruments. From hotlines to launch notifications, declaratory policies and no-first use pledges, to stabilizing measures such as the 1972 Soviet–US Anti-Ballistic Missile Treaty (ABM Treaty) and the 1987 Soviet–US Treaty on the Elimination of Intermediate-Range and Shorter-Range Missiles (INF Treaty), the web of risk-reduction measures helped ensure that the norm against the use of nuclear weapons stayed strong.[13]

Unfortunately, the suite of emerging technologies with implications for international peace and security has yet to develop such a safety net, predominantly because understanding of the impact of many of these technologies is still nascent. There are already interstate deliberations on autonomous weapons and

[10] On the problem of unpredictability of autonomous systems see also chapters 3 and 4 in this volume.

[11] President of Russia, 'Presidential address to the Federal Assembly', 1 Mar. 2018; and Mehta, A., 'LRS-B details emerge: major testing, risk reduction complete', *Defense News*, 2 Sep. 2015.

[12] E.g. International Committee of the Red Cross (ICRC), *Autonomous Weapons Systems: Technical Military, Legal and Humanitarian Aspects*, Expert meeting, 26–28 Mar. 2014, Geneva, Switzerland (ICRC: Geneva, 2014).

[13] See e.g. Borrie, J., Caughley, T. and Wan, W. (eds), *Understanding Nuclear Weapon Risks* (UN Institute for Disarmament Research: Geneva, 2017). See also Soviet–US Treaty on the Limitation of Anti-Ballistic Missile Systems (ABM Treaty), signed 26 May 1972, entered into force 3 Oct. 1972, not in force from 13 June 2002, *United Nations Treaty Series*, vol. 944 (1974), pp. 13–17; and Soviet–US Treaty on the Elimination of Intermediate-Range and Shorter-Range Missiles (INF Treaty), signed 8 Dec. 1987, entered into force 1 June 1988, *United Nations Treaty Series*, vol. 1657 (1991), pp. 4–167.

cybersecurity but, while some important normative understandings have been discussed, these deliberations have yet to produce anything binding and may not for some time.[14]

In parallel, the existing nuclear risk safety net is being eroded through the demise of bulwarks of the arms control framework such as the INF Treaty. Yet problems related to the nexus between nuclear risk and emerging technologies are barely permeating relevant nuclear forums. Risk-reduction conversations in forums such as the UN Disarmament Commission or the review cycle of the 1968 Non-Proliferation Treaty (NPT) are limited to references to cyber vulnerabilities.[15]

How to confront the new nuclear risks posed by technological innovation poses significant questions. Governance gaps are widening as technologies diffuse and converge in ways that further complicate the ability of states and international regimes to impose control.

II. Governing the risks posed by technological innovation

The only way to eliminate the risks posed by nuclear weapons is to eliminate nuclear weapons. However, the pursuit of a world free of nuclear weapons will require a spectrum of responses, including those needed to reduce the dangers posed by this nexus of nuclear weapons and technology. Some of these can be found in existing mechanisms, but some may require new approaches.

Bring together coalitions of non-traditional partners to explore the risks

Given the current levels of uncertainty, a first step could be to develop a better understanding of the risks. Doing so requires bringing together coalitions of non-traditional partners, from states and their militaries via intergovernmental organizations to civil society, academia and industry. The last of these—industry—is increasingly necessary as it includes the progenitors of much of the relevant technology. In this context, the UN, with its universal convening power, can play a significant role in providing the required platform to facilitate conversations and knowledge sharing.

Such non-traditional partners should also contribute to selected multilateral forums in an expert capacity. The Conference on Disarmament has shown how civil society, technical subject matter experts, industry and the research community can be incorporated into informal discussions, but this initiative needs

[14] E.g. Group of Government Experts of the Parties to the CCW Convention, CCW/GGE.1/2018/3 (note 8); and United Nations, General Assembly, Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, 22 July 2015, A/70/174. A similar group established by the UN General Assembly for the period of 2016–17 was unable to reach consensus on a final report.

[15] Preparatory Committee for the 2020 NPT Review Conference, 'Chair's factual summary (working paper)', NPT/CONF.2020/PC.II/WP.41, 16 May 2018, para. 28; and Treaty on the Non-Proliferation of Nuclear Weapons (Non-Proliferation Treaty, NPT), opened for signature 1 July 1968, entered into force 5 Mar. 1970.

to be boosted and replicated, including in the First Committee of the UN General Assembly, which addresses international security and disarmament issues. Incorporating practitioners into the multilateral system (e.g. in genuine expert bodies) will add an element of impartial technical expertise to deliberations on these matters. Exchange, interaction and cross-education are needed for effective policy development.

Such interaction would also provide opportunities to improve communication, establish a common vocabulary, build bridges, avoid redundancies, bust silos, boost reactivity and proactivity, and identify potential impacts of emerging technology with enough time to develop considered responses.

However, corporations, start-ups and universities working on emerging technologies do not need to wait for invitations to government- or UN-endorsed symposia.

Include technology-based risks as part of nuclear risk-reduction efforts

The current deteriorating security environment has given rise to growing support for urgent nuclear risk-reduction measures. The inclusion of technology-based risks must be a part of any deliberations. This should include the review process of the NPT, the cornerstone of the nuclear non-proliferation and disarmament regime, where discussions around threats, risks and opportunities are already taking place.[16] The NPT states parties should consider how to include measures to mitigate these new risks in any outcome of the 2020 NPT Review Conference.

Likewise, those bodies established to deal with the peace and security implications of emerging technologies, such as relevant UN groups of governmental experts, could consider this nexus.

Ideally, these various conversations would lead initially to the development of near-term politically binding confidence-building measures (e.g. enhanced transparency on how technologies are being incorporated into military and security doctrines) and agreements not to interfere with command-and-control structures or test or deploy destabilizing new capabilities.

Treaties have traditionally ruled the security domain, but they are at risk of becoming outpaced by technological change. Advances in science and technology, especially those with disruptive potential, will not wait for the long lead times needed for multilateral negotiations and ratifications.

Soft law and self-regulation for responsible innovation

When it comes to keeping ahead of technology, 'soft' law or self-regulating standards-based approaches might be valuable. These could include the development of codes of conduct or principles applicable to the development of new and potentially destabilizing technologies. Perhaps most importantly, they

[16] Preparatory Committee for the 2020 NPT Review Conference, NPT/CONF.2020/PC.II/WP.41 (note 15), para. 28.

should include a better understanding of the issue of foresight—that is, the ability to consider plausible ways in which a technology, system or feature might be used, not just how it was meant to be or should be used. Responsible innovation needs to be matched by forward-looking remediation. Measures to disseminate and share knowledge and to build strong, diverse and interdisciplinary communities of practice to cross-pollinate insights and experiences will help ensure that innovation is guided by risk assessment from the start.

Another near-term step should be to use the research community to examine the potentially beneficial impacts of technology on international peace and security. The creation of technical advisory bodies in multilateral deliberation bodies and international organizations, for example, would help policymakers to better leverage expertise and understand the benefits of new technologies (such as the Scientific Advisory Board of the Organisation for the Prohibition of Chemical Weapons and the various groups of scientific experts convened during negotiation of the 1996 Comprehensive Nuclear-Test-Ban Treaty). In the same way that these innovations can increase the quality and precision of weapons, so too can they enhance the set of tools available to facilitate their elimination.

III. Using machine learning and distributed ledger technologies to support compliance and verification regimes

Verifying nuclear disarmament

Verification and compliance are often cited as principle challenges to disarmament and non-proliferation efforts. Data-driven machine learning algorithms and AI-powered technologies and systems might enable new breakthroughs in compliance and verification regimes.

AI-powered technologies will allow states to consume and analyse vast quantities of information, while global networked communications will enable real-time transmission, enhancing confidence between partners. Although the adoption of blockchain and other distributed ledger technologies (DLTs) is still nascent, the variety of possible applications, including in the arms control and verification field, is also garnering interest.[17] DLTs could help to secure data and at the same time make it more transparent. Advances in image-recognition software coupled with the increasing availability and quality of satellite imagery could allow more actors to engage in verification activities.[18] This would effectively crowdsource what was once the domain of technologically sophisticated states.

Much more work is needed to examine the potential benefits that new technologies could have for nuclear disarmament verification.

[17] Vestergaard, C., 'Better than a floppy: the potential of distributed ledger technology for nuclear safeguards information management', Stanley Foundation Policy Analysis Brief, Oct. 2018; and Frazer, S. L. et al., *Exploratory Study on Potential Safeguards Applications for Shared Ledger Technology* (Pacific Northwest National Laboratory: Richland, WA, Feb. 2017).

[18] See e.g. Dorfman, Z., 'True detectives', *Middlebury Magazine*, spring 2018.

Preventing illicit procurement of weapons of mass destructions

Blockchain technology also holds potential, if current regulatory challenges are overcome, to contribute to improved measures to control acquisitions related to weapons of mass destruction programmes. These would help to prevent illicit procurement of goods and technologies by establishing robust supply-side control measures to ensure end-user verification and prevent export fraud.

Key information for controlled goods—such as export control classification numbers, end-users and other licensing information—could be included in the blockchain, which would be visible to all authorized parties. This would make it more difficult for unauthorized parties to fraudulently obtain and divert export-controlled goods.[19]

Monitoring nuclear tests

A promising research programme led by the University of California, Berkeley, applied machine learning to conduct seismic monitoring for nuclear tests. As the researchers note,

> Putting monitoring onto a sound probabilistic footing also facilitates further improvements such as continuous estimation of local noise conditions, travel time, and attenuation models without the need for ground-truth calibration experiments (controlled explosions). Moreover, it facilitates an open-source approach, whereby various expert groups can devise and test more refined and accurate model components and contribute them as modules in an open probabilistic architecture.[20]

Such an approach would complement existing efforts by the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) to use machine learning in its International Monitoring System.[21]

In another example, in 2018 the University of Tokyo launched a tool to predict the direction of radioactive material dispersion. Researchers used machine learning and computational methods to run meteorological simulations and analysis of data sets of near-surface wind conditions. The research demonstrated an average success rate of 85 per cent and was able to predict conditions up to 33 hours in advance.[22]

[19] Arnold, A., 'Blockchain: a new aid to nuclear export controls?', *Bulletin of the Atomic Scientists*, 19 Oct 2017.

[20] Arora, N. S., Russell, S. and Sudderth, E., 'NET-VISA: network processing vertically integrated seismic analysis', *Bulletin of the Seismological Society of America*, vol. 103, no. 2A (Apr. 2013), pp. 709–29, p. 728. See also Arora, N. S. et al., 'Global seismic monitoring: a Bayesian approach', *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence* (Association for the Advancement of Artificial Intelligence (AAAI): Menlo Park, CA, 2011), pp. 1533–36.

[21] Russell, S., Vaidya, S. and Le Bras, R., 'Machine learning for Comprehensive Nuclear-Test-Ban Treaty monitoring', *CTBTO Spectrum*, no. 14 (Apr. 2010).

[22] Yoshikane, T. and Yoshimura, K., 'Dispersion characteristics of radioactive materials estimated by wind patterns', *Scientific Reports*, vol. 8 (2018), article no. 9926, 2 July 2018.

IV. Conclusions

The challenges posed by the nuclear risk–technological innovation nexus, including governance challenges, are as daunting as they are urgent. Yet, as shown above, they should not be insurmountable if technological breakthroughs are balanced while properly harnessing relevant new technologies to ameliorate new and old nuclear risks.

What is needed now is for policymakers and developers to engage with a new cast of actors in order to explore the risks and opportunities and to mitigate the former while exploiting the latter.

Conclusions

15. Promises and perils of artificial intelligence for strategic stability and nuclear risk management: Euro-Atlantic perspectives

VINCENT BOULANIN

This edited volume is the first instalment of a trilogy that explores regional perspectives and trends related to the impact that recent advances in artificial intelligence (AI) could have on nuclear weapons and doctrines, strategic stability, and international security generally. It assembles the views of 14 experts from the Euro-Atlantic community who participated in a workshop on the topic organized by SIPRI in May 2018 in Stockholm.

This final chapter presents the key conclusions that can be drawn from this collection of essays. It first gathers the contributors' views on what the current AI renaissance offers and the risks that it brings (in section I). It then assesses what can be concluded about the impact of AI in the field of nuclear weapons and doctrines (in section II). Finally (in section III), it closes the volume by reviewing the options for dealing with the risks that accompany the conjunction of AI and nuclear weapons.

I. The promises and perils of the current AI renaissance

A nuanced understanding of the technology and associated risks is a precondition for an appropriate policy response

Getting the risk picture right

As a number of authors hint in their essays, it is easy to misconceive the opportunities and challenges posed by AI in the military domain in general and the nuclear domain in particular. The field of AI is going through a high-profile renaissance. There is a growing number of news articles, publications and public events that attempt to analyse the components of what the current success of the AI renaissance is—that is, a breakthrough in the area of machine learning that has unlocked major opportunities for the development of AI applications, such as autonomous systems. Nevertheless, there are enduring misconceptions about the possibilities and risks that AI could actually raise in the near term in the military sphere. As Frank Sauer notes (chapter 10), 'AI and machine learning are still simultaneously over- and underestimated by both the general public and policymakers'.

One of the reasons why they are being overestimated is the terminology, which triggers anthropomorphic representations. For Sauer, 'the "intelligence" component of the term AI evokes the wrong association, namely with human learning and human intelligence', which 'both differ significantly from the nature of AI and machine learning and what they are currently capable of'.

The present author also flags this terminological problem in the introduction to AI and machine learning (chapter 2): 'the way in which machine learning works has nothing to do with the way humans learn'. This discrepancy between human and machine intelligence and learning is clearly illustrated by the technical descriptions of how machine learning and autonomous systems work by Dimitri Scheftelowitsch (chapter 3) and Martin Hagström (chapter 4).

A second reason why the potential of AI is overestimated is the lack of awareness about the multiple technical and operational problems that slow down the adoption by the military of machine learning applications and autonomous systems. The first problem is the limitation of the technology itself. In his essay on the state of autonomous systems, Scheftelowitsch writes that, for many tasks and operating environment, 'the design of an autonomous system that can be used in practice is a considerable engineering, mathematical and political challenge. The reasons for this lie not necessarily in the autonomous decision-making as such, since it is often easy to provide an appropriate mathematical model, but in the various other, not necessarily technical, aspects of autonomy.' The state of the art, while impressive, still trails a long way behind the cultural perception of what autonomous systems ought to be able to do in a military context, namely operate safely and reliably in complex, uncertain and adversarial environment. A number of contributors underline that state autonomous systems are still too brittle, to reuse Michael Horowitz's words (chapter 9). Hagström also notes that, while advances in machine learning could improve the design of autonomous systems as well as offering qualitative improvement to a large variety of military applications, they also generate unique problems in terms of system predictability and reliability. He underlines that a characteristic of the models created by machine learning is that they are not transparent: their behaviour may therefore not be fully understandable and predictable to the humans who design and use them—which is problematic in a military context since 'From an operational point of view, the effects of a weapon system must be predictable to the commanding officer'. For Hagström, there is therefore an important gap to be filled between what machine learning can do at the experimental level and what it can be trusted to do when actually deployed; to be able to exploit the advances of machine learning, the military will first have to solve some complex testing and verification problems.

There are, in other words, many reasons not to exaggerate the impact of the current AI renaissance on the military. The contributions in this volume illustrate that AI could enable major qualitative improvement in many areas of warfare; however, foreseeable developments will be far more prosaic than the common representation of military AI in popular culture. Superintelligent AI or Terminator-like autonomous systems are not the type of technology that policymakers and the general public should worry about. Rather, they should be concerned by the fact that the military might underestimate or disregard the limitations of current AI technology.

The contributors present a fairly similar diagnosis of the limitations of current AI technology. Together, they point towards four factors.

1. *The brittleness of AI*. AI technology is limited to extremely narrow tasks. In Sauer's words, it may '[fail] spectacularly when confronted with a task [or environment] that differs slightly from what it was trained for'.

2. *The opacity and unpredictability of machine learning*. Machine learning-based systems may generate unexplainable outputs and unpredictable behaviours.

3. *The bias embedded in the systems*. AI systems, including systems trained by machine learning, may include human bias that can have detrimental effects, particularly when these systems are intended to support critical human decisions, such as a decision to use force.

4. *The vulnerability of AI systems*. As thoroughly demonstrated by Shahar Avin and S. M. Amadae (chapter 13), the integration of AI into military systems not only increases their potential vulnerability to cyberattack (by increasing the 'attack surface' as they put it) but also makes possible new types of attack, for instance spoofing involving data poisoning.

Take time to explore the technology-based risks before deployment

With these limitations in mind, nearly all the contributors warn that an immature adoption of the newest development in AI technology by the military, particularly in the context of nuclear weapon systems, could have dramatic consequences. They seem to agree that it would be prudent for states to devote time and resources to understanding these limitations and how they can be mitigated early in the research and development process.

However, as Page Stoutland notes (chapter 7), 'The potential performance benefits . . . may prove irresistible to developers and government sponsors'. This concern seems to be shared by other contributors, including Sauer and Justin Bronk. They note that some states might be ready to lower their system safety and reliability standards in order to maintain or develop their technological edge over their competitors. Speaking from a British and European perspective, Bronk concludes in his essay on unmanned combat aerial vehicles (UCAVs) and autonomous weapons (chapter 12) that 'Potential adversary powers (and most probably the USA) will not wait for West European powers to make up their mind before making lethal, highly autonomous aircraft'. For Bronk, European states, and the United Kingdom in particular, have a key role to play in influencing 'the construction of norms around these systems'.

The exploration of the risks and policy options needs to be inclusive

Beyond alliances

States need to not only develop and better understanding the opportunities and challenges posed by the military use of AI, particularly in the nuclear force-related context; they also need to discuss these with other states. Bronk hints that one way

to start this discussion is to engage with like-minded states. From the perspective of Western countries, that would mean engaging in a conversation within the North Atlantic Treaty Organization (NATO) or the European Union (EU).

While it would certainly be beneficial if Western countries could agree on the risk diagnosis or preferable policy response, that would not be enough. The discussion on the risks and norms that could govern the use of AI in the military sphere in general—and in the nuclear context in particular—also needs to take place between NATO member states, Russia and other nuclear-armed states such as China and India.

There are a number of ongoing arms control discussion tracks that provide opportunities for such discussions: the process on lethal autonomous weapon systems (LAWS) under the 1980 Convention on Certain Conventional Weapons (CCW Convention) for issues related to conventional use of military AI and, for nuclear-related concerns, the review process of the 1968 Non-Proliferation Treaty (NPT). Technology-based risk should also be part of bilateral nuclear risk-reduction deliberations, particularly between Russia and the United States.

Non-state actors have a part to play

States must also have a conversation with civil society organizations, academia and industry about the risks posed by AI technology and how these could be mitigated via various forms of governance. As Anja Kaspersen and Chris King rightly point out (chapter 14), these non-state actors have an essential role to play, and in particular industry 'as it includes the progenitors of much of the relevant technology'. Non-state actors can help states to better understand the speed and developmental trajectory of the technology. States in turn can help academia and industry become more aware of the security risks associated with the technologies that they research and develop.

As Kaspersen and King point out, 'the UN, with its universal convening power, can play a significant role in providing the required platform to facilitate conversations and knowledge sharing'. In fact, the United Nations is already allowing non-governmental actors to take part in informal discussions in a number of multilateral forums, such as the Conference on Disarmament or the CCW regime. For Kaspersen and King, 'this initiative needs to be boosted and replicated, including in the First Committee of the UN General Assembly'.

At the same time, it is useful to bear in mind that discussions conducted in the UN framework easily become politized—in some arms control forums the discussion has become so polarized that constructive dialogue has become difficult. Other avenues for multi-stakeholder discussions will therefore be needed. To enable a constructive discussion, it may be useful to find neutral venues and discussion tracks that are not already burdened by major political contentions.

II. The impact of AI on nuclear weapons and doctrines: What can be said so far

The contributors to this volume seem to share an understanding that the discussion on the challenges posed by AI in the field of nuclear weapons and doctrines can only be speculative. This is principally because of (*a*) the difficulty of predicting technological development in the area of AI and (*b*) the lack of open-source information on how military planners see the role of AI in future nuclear military modernization plans. They nevertheless agree on a number of points.

How AI could have an impact on nuclear weapons and doctrines

How AI could be used in nuclear weapon systems

First, the contributors unanimously note that recent advances in AI could be exploited in all aspect of the nuclear enterprise.

The present author (chapter 6) and Horowitz describe how machine learning could be used to boost the detection capabilities of extant early-warning systems and improve the possibility for human analysts to do a cross-analysis of intelligence, surveillance and reconnaissance (ISR) data. They also note that autonomous systems provide new possibilities for remote sensing operations, for instance in the context of anti-submarine warfare.

Several contributors raise the question of whether nuclear-armed states could use autonomous unmanned systems such as unmanned aerial vehicles (UAVs) or unmanned underwater vehicles (UUVs) for nuclear weapon delivery as an alternative to intercontinental ballistic missiles (ICBMs) as well as manned bombers and submarines. They agree that, while technically feasible, the nuclear-armed states seem to remain reluctant about that possibility. In reference to the US case, Bronk, Horowitz and the present author cite official statements that the USA does not see any role for unmanned bombers in nuclear weapon delivery.

Game changing technologies?

To the question of whether AI-driven developments in nuclear weapon systems will fundamentally transform the field of nuclear weapons and doctrines, the answer that seems to emerge from the various contributions is ‘no’, at least not yet—for three reasons.

First, nuclear weapon systems already rely extensively on AI technology and automation. The connection between AI and nuclear weapons is not new. The opportunities and challenges associated with the use of AI and automation in nuclear weapon systems have been known for decades. For instance, John Borrie (chapter 5) shows how, during the cold war, nuclear policymakers already ‘grappled with the questions of which assessment and decision-making roles are appropriate for delegation to machines and what is an appropriate level of delegation’. Hence, from this perspective, recent advances in machine learning

and autonomous systems reinforce, rather than alter, the existing applications of AI and automation.

Second, as discussed above, the newest advances in machine learning and autonomous systems have some technical limitations, which mean that their incorporation into nuclear weapon systems might take (a long) time. The field of nuclear weapon technology is generally known for its conservativeness: it has been historically slow at integrating the newest technological developments. States will need to crack difficult testing and evaluation problems to gain confidence that these developments can be certified for use.

Third, the real gaming-changing scenario would be a situation where AI technology enables a nuclear-armed state to credibly threaten another nuclear-armed state's second-strike capability. However, while there are a number of technologies under development that are specifically intended to do that (e.g. the USA's *Sea Hunter*, a prototype autonomous surface vehicle that could track down nuclear-armed submarines), these do not seem mature enough to represent a credible threat—yet.

The impact on strategic stability in the Euro-Atlantic context

The fact that the latest advances in AI are too immature to trigger a radical transformation in the field of nuclear strategy does not mean that they could not have a palpable effect on strategic stability, particularly in a Euro-Atlantic context.

As noted by several contributors, in particular Jean-Marc Rickli (chapter 11), the field of strategy is 'highly psychological'. The perception of an enemy's capability matters as much, if not more, than its actual capability. A nuclear-armed state could trigger destabilizing measures based only on the belief that its retaliatory capability could be defeated by another state's AI capabilities. That is where the inherent nature of AI technology becomes a major problem: the fact that it is software-based makes tangible evaluation of military capabilities difficult. Nuclear-armed states could therefore easily misperceive their adversaries' capabilities and intentions. For that reason, Rickli argues that 'the nuclear powers should consider, with the highest priority, communicating clearly and accurately about their AI capabilities'.

In the Euro-Atlantic context, one likely worrisome scenario would be a situation where Russia would try to offset the technological advantage of the USA in the conventional realm through further modernization of its nuclear arsenal. Petr Topychkanov (chapter 8) considers an even more destabilizing possibility: that Russia shifts from a defensive to an offensive nuclear doctrine.

To add nuance to this picture, a number of contributors underline that AI technology brings not only risks, but also opportunities for strategic stability. For instance, Horowitz explains that recent advances in AI could enhance stability as they would provide nuclear decision makers with better tools for crisis management. On the one hand, advances in machine learning could increase a military commander's ability to process ISR information and make critical decisions in a time-critical situation. Autonomous systems, on the other hand,

would provide new opportunities for remote sensing and for dissemination of information and orders in an environment where communication is otherwise denied.

Kaspersen and King also point out that recent advances in AI could support ongoing disarmament and arms control processes. They explain, for example, that AI could enable new breakthroughs in compliance and verification regimes: machine learning could help to monitor for nuclear tests or prevent illicit procurement for weapon of mass destruction programmes.

III. Options for dealing with the risks

Some solutions already exist

Many of the contributors make the case that AI systems and automated systems have been part of the nuclear deterrence architecture for decades, which means that the risks associated with their use are well known. Recent advances in machine learning and autonomy would, in that context, most likely exacerbate these risks rather than create new ones.

As Sauer points out, this also means that possible solutions already exist: there might be no need to reinvent the wheel. For Sauer, 'No-first-use doctrines and a lowering of the alert status of nuclear arsenals, for example, would buy valuable time during a crisis and allow for a closer evaluation of the signals received, and so prevent escalation due to miscalculation and misperception'. Rickli and Kaspersen and King also stress that a traditional approach to transparency and information sharing could help. More openness about nuclear modernization plans and more information sharing on AI-related developments via different dialogue tracks could mitigate the destabilizing potential of AI.

Development of new policies may be needed

The fact that some risk-mitigation measures already exist does not meant that states should shy away from exploring new policy options. This is particularly true at the multilateral level considering that bilateral discussions between Russia and the USA on nuclear disarmament and arms control issues have been deteriorating dramatically in recent years.

As Kaspersen and King note, a conversation on the need for and form of new risk-mitigation measures targeted at AI-related developments should take place in the framework of the NPT, which is 'the cornerstone of the nuclear non-proliferation and disarmament regime, where discussions around threats, risks and opportunities are already taking place'. In that regard, the upcoming NPT Review Conference in 2020 would provide great opportunity for states to engage on this topic.

The outcome of such discussions may not need to be the creation of a new treaty or new international laws. As Kaspersen and King stress, 'When it comes to keeping ahead of technology, "soft" law or self-regulating standards-

based approaches might be valuable'. These could include 'politically binding confidence-building measures (e.g. enhanced transparency on how technologies are being incorporated into military and security doctrines) and agreements not to interfere with command-and-control structures or test or deploy destabilizing new capabilities' or 'codes of conduct or principles applicable to the development of new and potentially destabilizing technologies'.

All in all, it appears from this collection of essays that there is a general consensus among experts from the Euro-Atlantic region on the impact that recent advances in AI could have on strategic stability and nuclear risk. It should be stressed, however, that their understanding of the challenges and how these can be dealt with seem to be influenced by the experience of the Soviet–US strategic relationship during the cold war. In East Asia and South Asia, geography and history have led the nuclear-armed states to develop a different understanding of what the key pillars of strategic stability are, including with regard to concepts such as deterrence, conflict prevention and resolution, and the factors that could cause a nuclear conflict. It is thus reasonable to assume that experts from these regions, whose views will be represented in the next two volumes of this series, will draw a different picture.[1]

[1] Saalman, L. (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. II, *East Asian Perspectives* (SIPRI: Stockholm, forthcoming); and Topychkanov, P. (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. III, *South Asian Perspectives* (SIPRI: Stockholm, forthcoming).

About the authors

S. M. Amadae (United States) is a university lecturer in politics at the University of Helsinki and also holds appointments as a research affiliate with the Science, Technology and Society (STS) programme of the Massachusetts Institute of Technology (MIT) and as associate professor of politics and international relations at Swansea University. Amadae researches deterrence, cooperation and norms in international relations and international political economy. She has recently published *Prisoners of Reason: Game Theory and Neoliberal Political Economy* (Cambridge University Press, 2016).

Shahar Avin (Israel) is a postdoctoral research associate at the Centre for the Study of Existential Risk of the University of Cambridge. Avin researches global catastrophic and existential risks, with a focus on artificial intelligence (AI) risks, using a mixture of methods that include modelling, foresight, expert elicitation and participatory scenarios. He was a co-lead author of *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation* (Feb. 2018), a multi-institution collaborative report.

John Borrie (New Zealand) is the research coordinator at the United Nations Institute for Disarmament Research (UNIDIR) and leads its programme on weapons of mass destruction and other strategic weapons. He is also an associate fellow at Chatham House. He has published on many aspects of arms control, disarmament, and other weapon-related processes and technologies. Borrie has a PhD from the University of Bradford.

Vincent Boulanin (France) is a senior researcher at SIPRI, where his work focuses on the challenges posed by the advances of autonomy in weapon systems and the military applications of AI more broadly. Before joining SIPRI in 2014, he completed a PhD in political science at the École des Hautes Études en Sciences Sociales, Paris. His recent publications include *Bio Plus X: Arms Control and the Convergence of Biology and Emerging Technology* (SIPRI, 2019, co-author).

Justin Bronk (United Kingdom) is a research fellow specializing in combat air power and technology at the Royal United Service Institute (RUSI). He is also editor of *RUSI Defence Systems*. Bronk has written on air power issues for the *RUSI Journal*, *RUSI Defence Systems*, RUSI Newsbrief, the *Journal of Strategic Studies* and *Air Power* and has contributing regularly to the international media. He is also a part-time doctoral candidate at King's College London.

Martin Hagström (Sweden) is deputy research director at the Swedish Defence Research Agency (Totalförsvarets forskningsinstitut, FOI). His areas of expertise are autonomous systems, aeronautics and unmanned vehicles, and his latest research includes work on autonomous weapon systems. Hagström is also the programme manager of the weapons and protection area and supports the weapons and protection research planning of the Swedish Armed Forces.

Michael C. Horowitz (United States) is professor of political science at the University of Pennsylvania and the associate director of its Perry World House. His books include *The Diffusion of Military Power: Causes and Consequences for International Politics* (Princeton University Press, 2010) and *Why Leaders Fight* (Cambridge University Press, 2015, co-author). He has published widely on the future of war, innovation, leadership and forecasting geopolitical events. His recent research focuses on how AI and robotics will shape global power. He received his PhD from Harvard University.

Anja Kaspersen (Norway) is director of the Geneva Branch of the UN Office for Disarmament Affairs (UNODA) and deputy secretary general of the Conference on Disarmament. Prior to taking up these positions in 2017, she was head of strategic engagement and new technologies at the International Committee of the Red Cross and senior director and Executive Committee member at the World Economic Forum.

Chris King (Australia) is deputy chief of the Weapons of Mass Destruction Branch and head of the Science and Technology Unit of UNODA. Before joining the UN, Chris served in the Australian Department of Foreign Affairs and Trade, including in India and Iraq. He was acting director of the Arms Control Section, responsible for Australian disarmament and non-proliferation policy, and an advisor to the Australian foreign minister, Kevin Rudd.

Jean-Marc Rickli (Switzerland) is the head of global risk and resilience at the Geneva Centre for Security Policy (GCSP). He is also a research fellow at King's College London and a senior advisor for the AI Initiative at the Future Society, Harvard Kennedy School of Government. He is a member of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems and the co-chair of the Partnership for Peace Consortium's working group on emerging security challenges. Rickli is co-author of *Surrogate Warfare: The Transformation of War in the Twenty-First Century* (Georgetown University Press, forthcoming).

Frank Sauer (Germany) is a senior researcher at Bundeswehr University Munich. His work focuses on the nexus between technology, society and security, in particular the application of AI and robotics in the military. His publications include *Atomic Anxiety: Deterrence, Taboo and the Non-Use of U.S. Nuclear Weapons* (Palgrave Macmillan, 2015) and 'Autonomous weapon systems and strategic stability', *Survival* (2017, co-author).

Dimitri Scheftelowitsch (Russia) is a postdoctoral researcher at TU Dortmund University in the areas of optimal decision-making under uncertainty and mathematical optimization. His current work focuses on mathematical models of uncertainty and corresponding optimization methods, continuing on from the subject of his PhD from TU Dortmund. Scheftelowitsch's recent publications include 'Computation of weighted sums of rewards for concurrent MDPs', *Mathematical Methods of Operations Research* (2019, co-author).

Page O. Stoutland (United States) is vice-president for scientific and technical affairs at the Nuclear Threat Initiative (NTI). He joined the NTI in 2010. He is responsible for NTI's scientific and technically related projects designed to strengthen nuclear security and reduce risks around the world. Current themes include working to strengthen cybersecurity for nuclear weapon systems and at nuclear facilities, promoting improvements in the security of nuclear materials through the NTI Nuclear Materials Security Index, and strengthening technical cooperation with China.

Petr Topychkanov (Russia) is a senior researcher with SIPRI's Nuclear Disarmament, Arms Control and Non-proliferation Programme. He works on issues related to the nuclear non-proliferation, disarmament, arms control and the impact of new technologies on strategic stability. From 2006 to 2017 he was a fellow in the Nonproliferation Program of the Carnegie Moscow Center. Topychkanov's recent publications include *Setting the Stage for Progress Towards Nuclear Disarmament* (SIPRI, 2018, co-author)