sipri

THE IMPACT OF ARTIFICIAL INTELLIGENCE ON STRATEGIC STABILITY AND NUCLEAR RISK

Volume II

East Asian Perspectives

EDITED BY LORA SAALMAN

October 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 II
East Asian Perspectives

EDITED BY LORA SAALMAN

October 2019

Contents

Preface

The post-cold war global strategic landscape is currently in an extended process of being redrawn. A number of different trends are in play here. 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 of the current United States' administration with, perhaps paradoxically, the international institutions and arrangements the USA had a big hand in creating. As a result, a binary Russian–US nuclear rivalry, a legacy of the old Russian–US confrontation, is being gradually augmented 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—let alone disarmament—is in doubt to an unprecedented degree.

On top of this comes the impact of new technological developments. 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 analysed. 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 on the global strategic landscape, and whether and how it might undermine international security. This edited volume on East Asian perspectives is the second major publication of this two-year research project. The authors are experts from China, Japan, South Korea, Russia and the USA. This volume was preceded by one on Euro-Atlantic perspectives and will be followed by one on South Asian perspectives, as well as a final report.

SIPRI commends this study to decision makers in the realms of arms control, defence and foreign affairs, to researchers and students in departments of politics, international relations and computer science, and to members of the general public who have a professional and personal interest in the subject.

Dan Smith
Director, SIPRI
Stockholm, October 2019

Acknowledgements

This edited volume is the second in a series generated by a two-year SIPRI research project that addresses two main questions related to the connection between artificial intelligence and nuclear weapons. First is the question of whether and to what extent machine learning and autonomy may become the focus of an arms race among nuclear-armed states. The second question is the impact that this may have on calculations of strategic stability and nuclear risk at the regional and transregional level.

The editor would like to express sincere gratitude to the Carnegie Corporation of New York (CCNY) for its generous support of the project. The editor is also indebted to all the experts who participated in the East Asia workshop that SIPRI and the China Institutes of Contemporary International Relations (CICIR) organized on the topic on 6–7 September 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 East Asia 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 SIPRI, CICIR, CCNY or any organization to which the authors are affiliated.

The editor also wishes to thank her SIPRI colleagues Sibylle Bauer, Vincent Boulanin, Petr Topychkanov and Su Fei for their constructive feedback and contributions. Finally, the editor would like to acknowledge the invaluable editorial work of SIPRI's Editorial Department.

Lora Saalman

Abbreviations

AI	Artificial intelligence
ATR	Automatic target recognition
C4ISR	Command, control, communications, computers, intelligence, surveillance and reconnaissance
CBM	Confidence-building measure
CCW	(Convention on) Certain Conventional Weapons
CFE	(Treaty on) Conventional Armed Forces in Europe
CPGS	Conventional Prompt Global Strike
DBN	Deep belief network
DOD	Department of Defense (of the USA)
FPI	Fond perspektivnykh issledovanii (Russian Foundation for Advanced Research Projects in the Defence Industry)
GGE	Group of governmental experts
ICBM	Intercontinental ballistic missile
ICT	Information and communications technology
ISR	Intelligence, surveillance and reconnaissance
KAIST	Korea Advanced Institute of Science and Technology
LAWS	Lethal autonomous weapon systems
MAD	Mutually assured destruction
MND	Ministry of National Defense (of South Korea)
OODA	Observe–orient–decide–act
NC3	Nuclear command, control and communications
NPT	Non-Proliferation Treaty
PLA	People's Liberation Army
R&D	Research and development
SSBN	Nuclear-powered ballistic missile submarine
START	Strategic Arms Reductions Treaty
UAV	Unmanned aerial vehicle
UCAV	Unmanned combat aerial vehicle
UN	United Nations
UUV	Unmanned underwater vehicle

Executive Summary

Artificial intelligence (AI) is not only undergoing a renaissance in its technical development, but is also starting to shape deterrence relations among nuclear-armed states. This is already evident in East Asia, where asymmetries of power and capability have long driven nuclear posture and weapon acquisition. Continuing this trend, integration of AI into military platforms has the potential to offer weaker nuclear-armed states the opportunity to reset imbalances in capabilities, while at the same time exacerbating concerns that stronger states may use AI to further solidify their dominance and to engage in more provocative actions. This paradox of perceptions, as it is playing out in East Asia, is fuelled by a series of national biases and assumptions that permeate decision-making. They are also likely to serve as the basis for AI algorithms that drive future conventional and nuclear platforms.

This volume, based on a workshop held in Beijing in September 2018, is the second of a series of three. They form part of a SIPRI project that explores regional perspectives and trends related to the impact that recent advances in AI could have on nuclear weapons and doctrines, as well as on strategic stability and nuclear risk. This volume assembles the perspectives of 13 experts from East Asia, Russia and the United States on why and how machine learning and autonomy may become the focus of an arms race among nuclear-armed states. It further explores how the adoption of these technologies may have an impact on their calculation of strategic stability and nuclear risk at the regional and transregional levels.

At the defensive level, integration of machine learning and autonomy into military platforms has a strong allure for countries with less capable early-warning systems, as well as smaller and weaker nuclear and conventional arsenals. East Asian experts highlight the advantages of machines undertaking decisions based on objective criteria to avoid the pitfalls of human error and to engage in faster anticipation, discrimination and response. For a country with concerns over a disarming first strike against its nuclear and conventional arsenals, as well as doubts over the reliability of countermeasures based on human response, AI-enabled systems provide a means to supplement and to even replace older military systems.

At the offensive level, platforms with longer endurance, such as unmanned underwater vehicles (UUVs), unmanned aerial vehicles (UAVs) and spaceplanes, provide resiliency and survivability. These two aims indicate why such vehicles are likely to be the AI-enabled platforms of choice for future nuclear delivery. However, some questions linger relating to development and deployment of such unmanned autonomous platforms, including over their very nature. Some countries in the region lack a clear differentiation as to whether certain UUVs and UAVs are to serve combat missions. Further, the line of distinction between unintentional and intentional collision remains unclear, along with the escalatory effect of such an incident when it involves at least one nuclear platform.

This crisis potential becomes even more difficult to gauge given that countries in East Asia are increasingly hedging on whether platform payloads will be conventional or nuclear, as with the DF-ZF hypersonic glide vehicle in the case of China or alleged short-range cruise missiles in the case of the Democratic People's Republic of Korea (DPRK, or North Korea). Even in the case of Russia, which has been more explicit as to the nuclear payload of its planned hypersonic vehicles and such UUVs as the Poseidon, the distinction between unintentional and intentional escalation remains a source of contention.

The development and deployment of AI-enhanced platforms have both been shaped by and have contributed to an interlocking series of national biases and assumptions that are driving AI integration and decision-making. As one example, the US contention that Russia's nuclear posture is predicated on first escalating a crisis in order to de-escalate it seemingly contributed to shifts in the 2018 US Nuclear Posture Review to advocate for low-yield nuclear platforms. Similarly, China's focus on fielding swarm-enhanced unmanned platforms in sea, air and space for surveillance and even engagement suggests a prevailing concern over the spread of US prompt and precise weaponry—such as Conventional Prompt Global Strike (CPGS)—that could result in decapitation of both its conventional and nuclear command and control and even arsenals.

The more autonomy and machine learning that is built into military platforms to address these intertwined concerns, the greater the importance of understanding regional perspectives that inform these systems and postures. If China is predicating decisions based on concerns over a US decapitating AI-enabled first strike, while the USA is reorienting its nuclear posture to address a perceived Russian 'escalate to de-escalate' nuclear strategy, then such factors must be taken into consideration when forming confidence-building measures.

Ultimately, AI is only an enabler or enhancement of often pre-existing systems. Signalling of intent among countries—rather than just technological advances—continues to be one of the more intractable issues. To mitigate miscalculation, it is essential to have a better understanding of the national biases and assumptions that are paramount drivers that contribute to AI-driven decision-making, nuclear posture and related technological advances. The workshops and the three volumes of this SIPRI project provide a space for experts from nuclear-armed and non-nuclear-armed states to engage in their own scenario building and analysis. This allows them to elucidate their perspectives to better address how AI is shaping nuclear risk in their respective regions and beyond.

Introduction

1. Introduction

LORA SAALMAN

In East Asia, there has been a heady burst of enthusiasm for and investment in the transformative power of artificial intelligence (AI) in both civilian and military modernization programmes. While a decade ago research on AI simply cited Western works dating to the early 1990s, AI integration has accelerated over the past few years with leadership pronouncements and national strategies demonstrating a desire to keep apace of these new technologies in the case of Japan and the Republic of Korea (South Korea) and even to dominate the field in the case of China and the Russian Federation. Even more so than nuclear weapons, the essays in this volume reveal a prevailing view that AI is the ultimate equalizer that can be capitalized on by a weaker state to bolster its conventional and nuclear forces. At the same time, a growing sense among some East Asian states that there will be a generation of AI 'haves' versus 'have-nots' has compelled them to decide not to allow themselves to fall behind in this newest arena of competition.

While AI may provide the information high-ground and dominance to which Chinese President Xi Jinping and Russian President Vladimir Putin refer in public statements, it also has strong psychological effects on countries that suspect that their advances are lagging. This traditional security dilemma has pushed countries to develop and introduce a host of AI technologies that exacerbate 'strategic time pressure'. This concept presupposes that the speeding up of decision-making on the battlefield—whether on land, at sea, in space or in cyberspace—compels military leaders to delegate greater command and control to machines. This volume reveals how this plays out across a spectrum of systems and platforms. In the nuclear arena, the potential that AI may enhance reconnaissance, speed, precision and manoeuvrability to the level that it renders a second-strike capability obsolete has contributed to an escalated adoption of prompt means of retaliation and even more offence-oriented postures that engage a spectrum of AI-enhanced options. Even before reaching the level of official doctrine, these technological advances have already begun to outpace national strategies.

In the case of countries such as China, Russia and the United States, which have placed AI acquisition and integration at the forefront of their military modernization, national AI strategic documents have emerged relatively late. Indeed, Russia's national strategy has still not been released to the public at the time of this writing. This poses some significant challenges in terms of determining intent, as well as developing confidence-building measures (CBMs) and controls. In an environment of distrust and alleged AI-enabled deep fakes, the authors of this volume recognize that these compromises may not even be viable in the near-, medium- or even long-term. Recognizing these lacunae, they describe the current state of this technology and its impact on strategic stability. In doing so, these experts offer a range of traditional and forward-looking CBMs to address this complex environment of emerging technologies and nuclear risk.

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.

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.[a]

[a] Podvig, P., 'The myth of strategic stability', *Bulletin of the Atomic Scientists*, 31 Oct. 2012.

Source: Boulanin, V., 'Introduction', ed. V. Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. I, *Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019), pp. 3–9, p. 4.

This collection of essays is based on the proceedings of a regional workshop—the second in a series of three—on mapping the impact of machine learning and autonomy on strategic stability and nuclear risk in East Asia, which SIPRI co-hosted with the China Institutes of Contemporary International Relations (CICIR) in Beijing in September 2018. This workshop assembled political, military, technical and academic experts from East Asia—China, Japan and South Korea—as well as experts from India, Pakistan, Russia and the USA. It consisted of a series of panel discussions that explored different aspects of the topic, as well as two break-out sessions in which smaller groups engaged in scenario-building exercises to analyse the risks that military applications of AI could pose to strategic stability and how to mitigate them. The terminology used here follows that used in the first volume, on the Euro-Atlantic region (see box 1.1).[1] A third volume, on South Asia, and a final report will follow.[2]

Overview

This volume is divided into two parts: part I covers the technologies and dynamics of AI and nuclear risk and part II explores the future of arms control and strategic stability with AI in East Asia. The volume concludes (in chapter 15) with a summary of the key conclusions drawn from the essays. The role of asymmetry

[1] Boulanin, V. (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. I, *Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019).

[2] Topychkanov, P. (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. III, *South Asian Perspectives* (SIPRI: Stockholm, forthcoming 2019); and Boulanin, V. et al., *Mapping the Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Final Report* (SIPRI: Stockholm, forthcoming 2019).

and signalling stand out as persistent leitmotifs throughout both the East Asia workshop and this volume.

AI in regional complexity

When applied to the nuclear domain, AI is seen within East Asia as having a dual role. On the one hand, it can be used to 'level the playing field' for countries that are weaker in conventional or nuclear terms. On the other hand, for countries that are dominant in these domains, some East Asian experts argue that AI may embolden these stronger powers to further engage in provocative behaviour, even a disarming conventional or nuclear first strike. This heightens concerns that a weaker adversary would lack the AI resources to anticipate and to engage in countermeasures, much less retaliation.

The dual-use, cross-domain nature of AI, combined with the already blurred lines between conventional and nuclear deterrence, demonstrates the difficulty of applying traditional definitions of strategic stability. As one East Asian participant at the workshop emphasized, every state in the region could be considered as equivalent to a nuclear power—even those without nuclear weapons—given their impact on regional strategic stability dynamics. Thus, while countries such as China, Russia and the USA set the bar in terms of applications of machine learning and autonomy relating to nuclear forces, Japan, the Democratic People's Republic of Korea (DPRK, or North Korea) and South Korea also have a significant role to play when it comes to nuclear risk.

This multiplicity feeds into the complex and broadened version of strategic stability posited by Cai Cuihong (China) in chapter 10. She argues that rather than defining the term as a narrow balance between two countries, it should instead reflect the breadth and multifaceted nature of contemporary strategic and AI dynamics, whether or not a country possesses nuclear weapons. Cai contends that nuclear weapons can only truly defend a country's core security interests, namely ensuring that the mainland will not face a large-scale attack from foreign enemies. To protect other national interests, she argues that the scope of strategic stability should be expanded to include conventional forces, along with technical, behavioural and institutional factors.

AI in cross-domain deterrence

As evident throughout the scenarios discussed during the East Asia workshop, pre-emption is not always about the nuclear dimension and is more likely to take place in other domains, such as cyberspace. To this end, Liu Yangyue (China) in chapter 3 argues that machine learning applications in cyberspace result in both an offensive and defensive dilemma that can expand the scale and dynamism of both detection and attack, thereby complicating traditional notions of deterrence. Li Xiang (China) in chapter 2 also explores this paradox. He explains that when a single cyber power dominates AI, it will enhance its offence–defence advantages in cyberspace and diminish stability, while two rivals possessing similar strengths in AI can engage in mutual deterrence.

Directing the focus towards specific platforms, the present author (USA) provides in chapters 4 and 8 case studies on AI-enabled cross-pollination between conventional and nuclear deterrence, exploring Chinese views on and development of unmanned vehicles transiting sea, air and space. In contrast with Russia, she notes that China has hedged to a greater extent on the intended payload of such platforms, particularly with its development of the DF-ZF hypersonic glide vehicle.[3] Still, she notes that China and Russia converge in their concerns over a 'false negative': their potential inability to anticipate an incoming disarming strike. Li Xiang (China) in chapter 2 echoes these concerns, maintaining that countries such as the USA may be able to use AI technology to improve its reconnaissance capabilities against China's mobile strategic missiles to discover the deployment rules, manoeuvring routes and launch site locations to eliminate the 'first-strike uncertainty' on which the Chinese nuclear deterrent is predicated.

Using a similar logic, the present author argues in chapter 8 that a country's assumptions about deficiencies in its early warning capabilities—combined with preoccupation with US advances in high-precision, stealthy and prompt systems—may encourage it to contemplate integration of machine learning, automation and autonomy into everything from launch-on-warning to neural networks that enhance manoeuvrability and precision guidance. She emphasizes that, while Chinese platforms have been traditionally developed in response to US military modernization and policy pronouncements, there is a marked difference when it comes to AI. Contrasted with Chinese publications of a decade ago that simply focused on foreign developments and countermeasures, she notes that China's technical communities have begun to develop their own models and strategic logic that in some domains mirror those of Russia.

AI in nuclear command and control

In terms of how this AI integration has an impact on nuclear force structures, experts in this volume explore the implications of deep learning algorithms on command and control. Cai in chapter 10 enumerates the main arenas in which AI excels to include cognition, prediction, decision-making and integrated solutions. Within this list, cognition refers to description of the world through the collection and interpretation of a wide range of data to feed predictive analysis of potential scenarios to better inform decision-making. By rooting decisions in pre-set goals, the idea is to provide an integrated solution for complex activities. When applied to nuclear and conventional forces, she argues that 'psychological anxiety' can lead to conflict escalation. Much of this stems from both perceived and real asymmetries in AI and nuclear capabilities among countries. Nishida Michiru (Japan) in chapter 14 further suggests that AI can shape many of these factors through enhanced command, control, communications, computers, intelligence, surveillance and reconnaissance (C4ISR) against enemy nuclear and conventional forces. He expresses concern that the dual-use nature of these systems makes

[3] Sayler, K. M., *Hypersonic Weapons: Background and Issues for Congress*, Congressional Research Service (CRS) Report for Congress R45811, (US Congress, CRS: Washington, DC, 11 July 2019); and Gault, M., 'Russia's new nuclear missiles squeeze response time', *Scientific American*, 27 Mar. 2019.

them particularly difficult to control, in part because they both enhance and undermine transparency and verification.

Vadim Kozyulin (Russia) in chapter 11 applies the above concepts and categories in examining how AI applications affect lethal autonomous weapon systems (LAWS). He highlights the shared concern that such platforms will select and engage their targets using unknown algorithms without meaningful human control or direct human supervision. While noting these common fears, he also explores the differences in national concerns over what AI means for military dominance in terms of missile defence, cyberattack, electronic suppression, and hypersonic and space weapons, which can purportedly enable a decapitating first strike.

In doing so, Kozyulin cites capabilities that are widely viewed within East Asia as tipping the scales to the advantage of countries such as the USA by enhancing prompt and stealthy attack, while suppressing the ability of the targeted country to engage in countermeasures and retaliation. This contention is comparable to the trends in China noted by Li in chapter 2 and Saalman in chapter 8. In addition to these asymmetries, Kozyulin notes that the radical reduction in the time required for C4ISR data analysis exacerbates the 'strategic time pressure' faced by militaries. This is salient in terms of how countries structure their nuclear and conventional forces, since the psychological and operational stress of being able to retaliate in a timely manner compels greater integration of automation and autonomy.

AI in military modernization

Contrasting Russia with the other nuclear powers, Vasily Kashin (Russia) in chapter 7 points out that it has been relatively late in its release of a comprehensive national programme for AI development. Nonetheless, he emphasizes that Russia has already advanced a series of AI-enabled platforms, including the Strategic Rocket Force's deployment of the Nerekhta autonomous combat vehicle, the RB-109A Bylina early-warning system and the Okhotnik unmanned combat aerial vehicle (UCAV). In particular, the Poseidon nuclear-powered unmanned underwater vehicle (UUV) that he cites promises to re-shape nuclear dynamics with its purported aim of enhancing Russia's second-strike capability.[4] In the light of such developments, Jiang Tianjiao (China) in chapter 9 echoes the concerns that some autonomous weapon platforms, including UUVs, increase the risk of accidental launch and nuclear war.

Hwang Il-Soon and Kim Ji-Sun (South Korea) in chapter 12 further highlight the challenges that such UUVs as the Poseidon may pose to strategic stability dynamics, as well as the very foundation of the 1968 Non-Proliferation Treaty (NPT).[5] They argue that, unlike nuclear platforms that engage in stage separation before detonation, these vehicles would explode both the nuclear warhead and the

[4] Peck, M., 'Russia has begun underwater tests of its Poseidon thermonuclear torpedo', *National Interest*, 19 May 2019.

[5] Treaty on the Non-Proliferation of Nuclear Weapons (Non-Proliferation Treaty, NPT), opened for signature 1 July 1968, entered into force 5 Mar. 1970.

nuclear reactor on-board, resulting in much greater and sustained contamination of the biosphere. When paired with the suggestion in the 2018 US Nuclear Posture Review that the USA could introduce low-yield submarine-launched ballistic missiles and low-yield submarine-launched cruise missiles, the complexity of regional dynamics is destined to grow with a sizeable impact on the posture of China and its neighbours.[6]

Moreover, other East Asian powers are also engaged in developments that may have second- and third-order effects as AI advances are integrated into national military and decision-support structures. This is critical for countries such as Japan, North Korea and South Korea that are on the front lines of some of the most potentially destabilizing developments in the conventional and nuclear military fields. Hwang Ji-Hwan (South Korea) in chapter 5 and Su Fei (China) in chapter 6 detail South Korea's research on AI-based command systems, aviation training systems and object-tracking techniques, as well as work on the Exobrain and ADAMs projects for potential enhancement of C4ISR, the Dronebot Jeontudan military unit, omni-directional movement interactive software technology for virtual combat exercises, and navigation algorithms for large-scale UUVs.

These authors compare such developments with those of North Korea. While relatively nascent, North Korea has pursued its own AI technology-based version of the game go, which it has also sought to apply in other areas. Reviewing the work of Kim Il Sung University and the AI Institute of the Korea Computer Center, Hwang suggests that the country has advanced the Ryongnamsan 5.1 speech-recognition system, while exploring such machine learning topics as audio classification and a fingerprint and facial recognition system. Su furthers this analysis by detailing North Korean applications of artificial neural networks in both autonomous robotics and cyber operations. These developments suggest that, in addition to kinetic platforms that result in physical damage, North Korea may be positioning itself to engage in more operations potentially enabled by deep fakes and data poisoning.

When it comes to Japan, Arie Koichi chapter 13 suggests that US extended nuclear deterrence for both Japan and South Korea could be undermined by increased applications of AI in conventional and nuclear forces. He thus argues for the inclusion of Japan and South Korea in consultative mechanisms on these technologies. Nevertheless, in line with Nishida in chapter 14 and other Japanese experts at the SIPRI workshops, their focus tends to remain trained on overall arms control dynamics and trilateral relations among China, Russia and the USA. Due to the military and official positions of these experts, the tendency to focus on these three parties and overall nuclear dynamics is understandable. Yet whether this attention also indicates Japanese concerns over the current state of US extended deterrence commitments, the extent of Japanese transparency on integration of machine learning and autonomy, or the level of development of

[6] US Department of Defense (DOD), *Nuclear Posture Review* (DOD: Washington, DC, Feb. 2018); Schneider, M. B., 'Escalate to de-escalate', *Proceedings* (US Naval Institute), vol. 143, no. 2 (Feb. 2017); and Oliker, O. and Baklitskiy, A., 'The Nuclear Posture Review and Russian "de-escalation": a dangerous solution to a nonexistent problem', *War on the Rocks*, 20 Feb. 2018.

Japan's own military AI integration is unclear. Thus, Japan's overall approach to AI and the nuclear domain remains an area for further research and inquiry.

AI in arms control

When it comes to AI-related CBMs and arms control, experts from Japan, Russia and South Korea offer the most concrete details. Nishida in chapter 14 provides an overview of the evolution of arms control and CBMs and how AI technology may fit into weapon- and behaviour-focused controls. To be successful in implementation, he stresses that the object targeted by arms control must have a clear definition to make it distinguishable from other non-controlled weapons. Moreover, the control measure needs to be verifiable. However, due to the highly dual-use nature of AI, he recognizes the difficulty of enforcing these demands. Thus, when placing applications of AI along a spectrum from offensive to defensive, he makes a compelling case as to why more defensive applications of AI, such as for early warning, show more promise than offensive applications in nuclear forces or attacks on nuclear command and control. This theoretical discussion further compliments chapter 12, where Hwang and Kim argue for nuclear-powered autonomous nuclear weapon vehicles to be integrated into the agenda of the preparatory committee for the 2020 NPT review conference. In effect, this platform could serve as the decisive test for the resiliency and feasibility of CBMs discussed by Nishida.

Kozyulin in chapter 11 offers further suggestions for models that could be applied in developing arms control relating to the integration of machine learning and autonomy into systems that have an impact on nuclear risk. Among the bodies and documents, he lists the work of the fifth review conference of the 1980 Convention on Certain Conventional Weapons (CCW Convention) on LAWS and the Tallinn Manual on the International Law Applicable to Cyber Warfare as examples of how to define and regulate disruptive technologies.[7] He also reviews such agreements as the Vienna Document 2011 on Confidence- and Security-Building Measures, suggesting such updates as including remotely operated or autonomous UCAVs in their coverage.[8] He further notes that such measures could be suitable for future application in East Asia. However, recognizing the currently tense international environment, he concludes by setting his sights on more near-term CBMs. In doing so, Kozyulin, much like Nishida in chapter 14, suggests that information sharing and other controls may occur along a transparency continuum. This concept of a spectrum along which experts can evaluate and address the impact of AI on strategic stability and nuclear risk serves as an apt entrée into exploring the East Asia workshop and this volume.

[7] Schmitt, M. N. (ed.), *Tallinn Manual on the International Law Applicable to Cyber Operations* (Cambridge University Press: Cambridge, 2013); Schmitt, M. N. (ed.), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press: Cambridge, 2017); 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, or 'Inhumane Weapons' Convention), opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983.

[8] Vienna Document 2011 on Confidence- and Security-Building Measures (Vienna Document 2011), adopted 30 Nov. 2011, entered into force 1 Dec. 2011.

Part I. The technologies and dynamics of artificial intelligence and nuclear risk

This part seeks to ground the reader's understanding of the technologies and platforms that are shaping the future of artificial intelligence (AI) and its integration into military modernization programmes throughout East Asia. From kinetic delivery and defence platforms to non-kinetic cyber networks that facilitate communications, intelligence and command-and-control operations, these seven essays provide an overview of some of the key technological trends and dynamics among such regional actors as China, North Korea, South Korea, Russia and the United States. They illustrate some of the underlying concerns over asymmetry and signalling that are endemic to the region.

The first three essays explore the manner in which AI has an impact on technological developments with both kinetic and non-kinetic consequences for deterrence. Li Xiang (in chapter 2) analyses how AI technology is altering strategic weapons, reconnaissance, missile defence, cyberspace, lethal autonomous weapon systems (LAWS), and nuclear command, control and communications. In doing so, he creates a foundation for understanding the pervasive reach of AI as an enabling technology. Liu Yangyue then details (in chapter 3) the dual implications of advances in machine learning for both strengthening and defending against cyber intrusions and attacks. The present author (in chapter 4) focuses this deterrence discussion on Chinese publications on hypersonic glide vehicles to discuss how manoeuvrability, targeting and autonomy enhancements may reflect a shift towards a more offence-oriented posture. All three essays indicate how these developments are both driven by and have an impact on perceptions, suggesting that deterrence depends a great deal on whether these capabilities are in the hands of the weaker or stronger nuclear-armed state.

The second set of essays delves into the integration of machine learning in North Korea and South Korea. Hwang Ji-Hwan (in chapter 5) provides an overview of some of the nascent AI programmes in both countries. He notes the ongoing lack of transparency and how it creates a black box in terms of understanding the military intent behind and the concrete advances stemming from this research. Nonetheless, he posits that both countries' sizeable cyber advances indicate that integration of machine learning into military applications is in the offing. Su Fei (in chapter 6) follows this foundation by using open source information to discuss how North Korean and South Korean systems may shape current and future applications of unmanned systems and cyberwarfare. She explores technological advances made by the South Korean Army's AI Research and Development Center, the Korea Advanced Institute of Science and Technology (KAIST), Hanwha Systems, and the Research Center for the Convergence of National Defense and Artificial Intelligence. She follows this with a discussion of North Korean work on artificial neural networks in cyberspace and autonomous mobile robotics based at Kim Il Sung University.

The last group of essays in part I examines Chinese and Russian military modernization through the prism of autonomous platforms. Vasily Kashin (in chapter 7) discusses the rationale behind such Russian strategic platforms as the Poseidon autonomous underwater nuclear-delivery platform. He argues that, while Russia's national programme for AI has not yet been articulated, its application is already progressing and is likely to include greater future collaboration with China. The present author (in chapter 8) then compliments this assessment with her own research on Chinese platforms and how China and Russia share some fundamental concerns about US stealth, high-precision and prompt platforms that threaten their conventional and nuclear deterrents. She notes that these shared perceptions are among the factors driving China towards deterrence based on avoiding 'false negatives', which may compel it to potentially adopt greater integration of automation and autonomy into its nuclear forces.

LORA SAALMAN

2. Artificial intelligence and its impact on weaponization and arms control

LI XIANG*

In recent years, the maturity of military artificial intelligence (AI) has advanced rapidly. Open source information indicates that the United States, Russia, the United Kingdom, France, China, Japan and the Republic of Korea (South Korea), among other countries, are engaged in such developments. Among the weapons and equipment that have been deployed in various countries, unmanned aerial vehicles (UAVs), unmanned underwater vehicles (UUVs), unmanned surface vessels, battlefield robots and other platforms have been put into use. AI technology based on big data, cloud computing, neural network-based deep learning, computer vision, intelligent robots, natural language processing and speech is of great military value. It can play a key role in intelligence monitoring and reconnaissance, target recognition, communication and navigation, automated command and control, firepower strikes, and cyber-electromagnetic countermeasures. Because AI enables improvement in the operational effectiveness of weapons and equipment, the future format of warfare is likely to be altered.

This essay first analyses (in section I) how AI technology is shaping strategic weaponry in terms of strategic reconnaissance, missile defence, and nuclear command, control and communications (NC3). It then (in sections II and III) reviews these AI-related developments in the realm of cyberspace and lethal autonomous weapon systems (LAWS). It concludes (in section IV) by suggesting means of mitigating the negative impact of some of these developments on traditional arms control.

I. Strategic weapons

Missile defence and strategic reconnaissance

AI can improve the effectiveness of missile defence and enhance target recognition, trajectory calculation and judgement of damage effects. It thereby improves the ability of countries with missile defence systems to offset their opponents' nuclear retaliation.

Further, the application of AI technology has a marked impact on strategic stability among nuclear powers and may undermine the mutual vulnerability and strategic stability of nuclear-armed states of unequal power. Because AI boasts strong capabilities in image and pattern recognition, similar to facial recognition, its ability to recognize still images is strong and it will greatly enhance the effectiveness of strategic reconnaissance. By improving the ability to interpret

* The views expressed in this essay are those of the author and do not necessarily reflect those of any organization to which he is affiliated. It was translated from Chinese to English by the volume editor, Lora Saalman.

satellite images and the deep reconnaissance of long-distance UAVs, the country that possesses nuclear superiority will be able to gain further intelligence on the basic characteristics and procedures of its opponent's nuclear force deployment and movements. This will enhance a country's confidence in its ability to disarm its opponent's nuclear weapons with a first strike. When there is a large gap between the nuclear forces of the two countries, this situation will make a pre-emptive strike by the more powerful one more advantageous, thus weakening the strategic stability of the two countries during a crisis.

At sea, AI technology can be used to improve the capability to collect and process a submarine's sound signature. In anti-submarine warfare, UUVs can also be used to conduct close-range reconnaissance, which can strengthen the capacity to detect and recognize enemy nuclear-powered ballistic missile submarines (SSBNs). On the one hand, this reduces the survivability of the enemy SSBNs, thereby weakening the effectiveness of nuclear deterrence and reducing strategic stability.[1] On the other hand, this AI technology can reduce the probability of accidental nuclear war: by improving the ability to identify SSBNs, a country can avoid accidentally hitting an SSBN as part of conventional anti-submarine warfare.

Facing the huge quantitative and qualitative advantages of the USA in nuclear weapons, China has maintained a slim and effective retaliatory nuclear force based on mobile strategic missiles with relatively strong concealment, manoeuvrability and survivability. This force is predicated on increasing the opponent's 'first-strike uncertainty' (第一次打击的不确定性) to ensure credible deterrence.[2] However, once the USA is able to effectively employ AI technology to improve its reconnaissance capabilities against China's mobile strategic missiles—enabling discovery of deployment rules, manoeuvring routes and launch site locations—this 'first strike uncertainty' would be eliminated.

The USA would thereby gain the advantage of being able to decapitate China's nuclear arsenal. To overcome concerns over this potential, China would have to increase the alert level of its nuclear weapons to ensure the credibility of its deterrent. This would lead to a state in which both sides tend toward pre-emption, weakening the strategic stability of the two countries and increasing the risk of nuclear conflict. In the case of sea-based nuclear deterrence, the employment of AI-based detection technology and offensive UUVs may also have similar effects. In particular, due to the difficulty of underwater communication, it would be difficult for these weapon platforms to receive onshore command and control signals in a timely manner. This would make it difficult to control, much less recall such platforms when engaged in operations.

At the current state of technological development, the USA has already begun to use AI in strategic reconnaissance and strategic anti-submarine search. In

[1] See also e.g. Rickli, J.-M., 'The destabilizing prospects of artificial intelligence for nuclear strategy, deterrence and stability', ed. V. Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. I, *Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019), pp. 91–98, p. 94.

[2] Wu, R., 'Certainty of uncertainty: nuclear strategy with Chinese characteristics', *Journal of Strategic Studies*, vol. 36, no. 4 (June 2013), pp. 579–614.

2017 researchers at the University of Missouri, USA, published a deep learning model for remote sensing satellite image recognition, which was trained using 2200 satellite images of surface-to-air missile positions.[3] It was able to identify in 42 minutes Chinese defensive air-to-air missile positions that would normally take a human nearly 60 hours to identify visually, with an accuracy rate of approximately 90 per cent. While this procedure faced many difficulties in the identification of camouflaged positions, the USA has recognized the potential advantages of these technologies. Thus, by December 2018 the US Office of Naval Research requested a white paper to study analytical research on the relationship between physical oceanographic changes and sound transmission, including field operations to collect relevant data sets.[4] It also engaged in analysis of large oceanographic and acoustic data sets, which factored in the development and use of AI and machine learning techniques.

Nuclear command, control and communications

The application of AI technology in the command, control and communications of nuclear weapons raises the question of whether the decision to use nuclear weapons will be determined by humans or by machines. A fully automated nuclear command-and-control system may increase the risk of accidental nuclear war. The commonly cited example is the Soviet Dead Hand system, which would automatically launch its nuclear missiles if its seismic, light, radioactive and pressure sensors detected that a nuclear weapon attack was underway.[5] In 1983, a false alarm by the Soviet early-warning satellite brought about just such an unexpected nuclear crisis. This crisis—known as the Petrov incident—was resolved due to the intervention of human judgment.[6] At the same time, there remains the issue of unmanned anti-submarine equipment. If the target of anti-submarine warfare is a non-strategic submarine, but misjudgement leads to an attack on a SSBN, this could also trigger an unexpected nuclear war.

Thus, while human judgment is not necessarily always reliable, in order to cope with various uncertainties and to assume responsibility for the use of nuclear weapons, the ultimate control of the nuclear command-and-control system still needs to be in the hands of human beings. In this regard, General John Hyten, head of US Strategic Command, has also said that once a computer system using AI is fully operational, the US Department of Defense (DOD) should consider taking security precautions to ensure that humans, not machines, control the decision on whether or not to use nuclear weapons.[7]

[3] Marcum, R. A., Davis, C. H., Scott, G. J. and Nivin, T. W., 'Rapid broad area search and detection of Chinese surface-to-air missile sites using deep convolutional neural networks', *Journal of Applied Remote Sensing*, vol. 11, no. 4 (Oct–Dec 2017).

[4] Tucker, P., 'How AI will transform anti-submarine warfare', Defense One, 1 July 2019.

[5] Borrie, J., 'Cold war lessons for automation in nuclear weapon systems', ed. Boulanin (W 1), pp. 41–52.

[6] Topychkanov, P., 'Autonomy in Russian nuclear forces', ed. Boulanin (note 1), pp. 68–75.

[7] Stewart, P., 'Deep in the Pentagon, a secret AI program to find hidden nuclear missiles', Reuters, 5 June 2018.

II. Cyberspace

AI contributes to enhancing cyber-deterrence, which can mean deterrence of an attack by cyber means or deterrence of a cyberattack.[8] From a technological perspective, the effectiveness of cyber-deterrence depends on the ability to identify the source of the cyberattack and the ability to engage in cyber-offence and cyber-defence. In other words, it requires the ability to discover intruders and to block or retaliate against intrusions. When it comes to the offence–defence paradigm in cyberspace, strategic-level cyberattacks rely on the mastery of vulnerabilities, as well as long-term penetration and sophisticated planning. Cyber-defence depends on detecting security threats, while making a timely response and rapid recovery after an attack. When coordinating complex attacks, human factors are of greater importance, since they often mark the greatest weakness and engage in multidimensional planning. When coordinating defence, AI's situational awareness, fast calculation and data-processing capabilities are more efficient than manual ones.[9] Therefore, AI is well-suited to play a greater role in network defence.

To this end, with the assistance of AI technology that facilitates the collection and processing of large amounts of historical cyberattack data, the defending party is better able to anticipate the attackers' means and rules of attack. AI can (*a*) improve the ability of the defender to identify the attack surface and the sources of attack, (*b*) reduce the anonymity of the attacker, (*c*) expose the attacker's activities, (*d*) provide warning and even threaten retaliation, and (*e*) offer cost-effective deterrence and retaliation. Furthermore, through training on cyberthreat perception and response, AI technology can detect and block devices that have been attacked and prevent the installation and operation of malware and files. In doing so, it can improve the operational efficiency of security operation centres, quantify network security risks, monitor network traffic anomalies, enhance cyber-defence and cyber-recovery capabilities, and improve deterrence by denial.

When a single cyber power has this capability, it will further gain situational awareness and offence–defence advantages in cyberspace, resulting in low strategic stability. When a rival to the major power has this capability, it will probably be able to engage in mutual deterrence and thereby improve strategic stability within cyberspace.

[8] Brantley, A. F., 'The cyber deterrence problem', eds T. Minárik, R. Jakschis and L. Lindström, *CyCon X: Maximising Effects*, 10th International Conference on Cyber Conflict (NATO Cooperative Cyber Defence Centre of Excellence: Tallinn, 2018); Cimbala, S. J., 'Nuclear deterrence in Cyber-ia: challenges and controversies', *Air & Space Power Journal*, vol. 30, no. 3 (fall 2016), pp. 54–63; and Raitasalo, J., 'Cyber deterrence is an oxymoron for years to come', *National Interest*, 20 Nov. 2018. See also chapter 3 in this volume.

[9] Naikal, N., *Towards Autonomous Situation Awareness*, Technical Report No. UCB/EECS-2014-124 (University of California at Berkeley Electrical Engineering and Computer Sciences: Berkeley, CA, 21 May 2014).

III. Lethal autonomous weapon systems

In terms of target identification, LAWS raise the question of whether such systems would be able to accurately distinguish between military and non-military targets.[10] This raises the additional issue of whether the systems would be able to effectively identify people's behaviour patterns, as with surrendering enemies. Further, such platforms may result in collateral damage. During the US administration of President Barack Obama, manually remote-controlled unmanned combat aerial vehicles (UCAVs) used in counterterrorism operations in Afghanistan and Pakistan caused civilian casualties.[11] This history raises the issue of who will ultimately bear the responsibility for the collateral damage of future AI weapons. Finally, LAWS are vulnerable to proliferation.[12] In the absence of a trade control system for such weapons, their cost-effectiveness and savings of manpower mean that they may be subject to greater use and abuse among countries and even non-state actors.

The above problems require that LAWS—including UCAVs, smart ammunition and combat robots among others—should have more accurate target recognition and greater precision in their killing ability. Russia has even deployed combat robots on the Syrian battlefield.[13] Such developments require a higher technical threshold. The collateral damage caused by LAWS is destined to become the focus of future arms control. Such arms control measures should be based on the technical level and strategic stability considerations of these platforms. Weapons that fail to meet the relevant technical standards should be subject to restrictions, and countries should be constrained in their production and deployment.

IV. Conclusions

Strategic weapons

In order to integrate AI-related arms control into strategic weapons, confidence-building measures serve as a good first step. As far as China and the USA are concerned, they must develop a mechanism that allows them to better understand each other's intentions in the development and application of AI-enabled strategic weapons through mutual information exchanges. Moreover, both should recognize their mutual strategic vulnerability to ensure that the USA does not use its advantages in strategic reconnaissance to initiate pre-emptive strikes in times of crisis. Nonetheless, following the release of the 2018 US Nuclear Posture

[10] Gerry, R., 'Making laws for LAWS: the legality of lethal autonomous weapon systems', Victoria University of Wellington, 2016.

[11] Bergen, P. et al., 'Drone strikes: Pakistan', New America, accessed 21 Aug. 2019; and Jaeger, D. A. and Siddique, Z., 'Are drone strikes effective in Afghanistan and Pakistan? On the dynamics of violence between the United States and the Taliban', *CESifo Economic Studies*, vol. 64, no. 4 (Dec. 2018), pp. 667–97.

[12] Chartoff, 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).

[13] Atherton, K. D., 'Russia eager to prove recent conflicts improved its robots', 27 June 2019, C4ISRNet; and Roblin, S., 'Russia's Uran-9 robot tank went to war in Syria (it didn't go very well)', *National Interest*, 6 Jan. 2019.

Review—with its strongly negative focus on China and Russia combined with its advocacy for nuclear modernization and low-yield nuclear weapon platforms—the potential for confidence-building measures faces serious challenges.[14]

In the light of such limits, it is crucial for the weaker nuclear-armed state in the strategic stability relationship to make a technical and tactical assessment of whether to employ big data and AI to strengthen its strategic reconnaissance. In doing so, the survivability of its mobile strategic missiles could be enhanced through more effective AI-enabled measures, such as concealment, deception, interference and camouflage. In terms of big data and AI techniques, this could include improvement to the design and routes of missile manoeuvres to counter satellite reconnaissance patterns. On SSBNs, this could involve employment of AI technology to improve quietness and to expand the scope of underwater activities. For missile defence, AI technology could be used to improve the penetration capability of missiles and other delivery systems, such as carrying more powerful decoys.

Cyberspace

It is crucial to place particular emphasis on the application of AI technology in cyberattack detection, tracking and recovery capabilities, while improving overall cyber-defence and cyber-deterrence. Countries that have common interests in avoiding cyberattacks can use AI technology to enhance the transparency of cyberspace and to mitigate anonymous attacks. When the two sides establish strategic stability on the basis of effective cyber-deterrence, this could be similar to nuclear arms control.

Such measures could include the pursuit of arms control consensus on such measures as reciprocal actions and unilateral commitments to not engage in first strike or attacks on each other's critical infrastructure. This would promote the formation of a cyberspace arms control system. To this end, China and Russia have signed an agreement on safeguarding international information security, which includes a pledge not to engage in cyberattacks against each other.[15] China and the USA also pledged in 2015 not to attack each other's critical infrastructure.[16]

Lethal autonomous weapon systems

Countries that are developing LAWS should consider the humanitarian issues arising from the use of such weapons and continue to promote their regulation in the framework of the 1980 Convention on Certain Conventional Weapons (CCW

[14] US Department of Defense (DOD), *Nuclear Posture Review* (DOD: Washington, DC, Feb. 2018), pp. 54–55.

[15] Chinese Ministry of Foreign Affairs, '中俄签署国际信息安全合作协定' [China and Russia sign international information security cooperation agreement', 12 May 2015.

[16] White House, Office of the Press Secretary, 'President Xi Jinping's state visit to the United States', Fact sheet, 25 Sep. 2015.

Convention).[17] Additionally, they should further engage in efforts to discuss and define a scope for LAWS that takes into consideration rules of engagement, performance and technical indicators to develop relevant standards and specifications for norms.

The major AI powers should be aware of the potential for LAWS to be included in arms control in the future. As such, they should take into consideration arms control factors when developing, producing and deploying related weapons, so that they are better prepared to participate in the arms control process when it reaches greater maturity.

[17] 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, or 'Inhumane Weapons' Convention), opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983.

3. The role of artificial intelligence in cyber-deterrence

LIU YANGYUE*

Developments in artificial intelligence (AI) have progressed rapidly in recent years. Through advances in machine learning and other applications, a wide range of new applications have emerged in cybersecurity and cyber-deterrence.[1] This essay seeks to shed some light on these concepts, as they relate to AI integration. It first briefly reviews the potential impact of machine learning on cyber-deterrence (in section I). It then describes the conditions for effective cyber-deterrence (in section II) and some consequent problems for cyber-deterrence (in section III). The essay concludes (in section IV) by advocating for the importance of greater interaction on these issues before the technologies outpace the ability of countries to reach greater understanding and consensus.

I. The impact of machine learning

Machine learning is designed to train a computer to complete a certain task on its own. It is expected to change the cyberspace landscape in several ways.

First, new algorithms using machine learning are more adaptive, offering enhanced dynamism. Because cybersecurity risks evolve quickly over time, new generations of malware and cyberattacks are difficult to detect with traditional cybersecurity protocols. Machine learning overcomes this weakness by allowing cybersecurity systems to use pre-existing cyberattack data to respond to similar attacks.

Second, machine learning reduces the need for human labour in cybersecurity interactions, both in terms of offence and defence. A typical example of this would be spear-phishing, which tricks a specific individual or organization into leaking confidential information. Traditional methods of spear-phishing are often of limited scope. Effective intrusion requires a large amount of research on the potential target. Moreover, it is difficult, if not unfeasible, to attack multiple targets simultaneously. Yet, with the help of machine learning, automation of spear-phishing may be possible.

The third area in which machine learning may make a difference is attribution. With more powerful learning capabilities, these algorithms are better able to

[1] Brantley, A. F., 'The cyber deterrence problem', eds T. Minárik, R. Jakschis and L. Lindström, *CyCon X: Maximising Effects*, 10th International Conference on Cyber Conflict (NATO Cooperative Cyber Defence Centre of Excellence: Tallinn, 2018); Cimbala, S. J., 'Nuclear deterrence in Cyber-ia: challenges and controversies', *Air & Space Power Journal*, vol. 30, no. 3 (fall 2016), pp. 54–63; and Raitasalo, J., 'Cyber deterrence is an oxymoron for years to come', *National Interest*, 20 Nov. 2018.

* The views expressed in this essay are those of the author and do not necessarily reflect those of any organization to which he is affiliated.

locate critical evidence to reveal the real identity of an attacker, for example if certain code fragments mimic existing malware structures.

II. Conditions for cyber-deterrence

Despite its extensive use, the concept of cyber-deterrence remains unclear and under debate.[2] There is disagreement over whether cyberattacks could be effectively deterred and even about whether the notion of deterrence is meaningful in cyberspace. However, according to those who are in favour of cyber-deterrence, it is feasible when certain conditions are met. For example, cyber-deterrence may work when it makes the cost of cyberattack exceptionally high. This can be achieved either by enhancing cyber-defence, thus enabling 'deterrence by denial', or by making retaliation credible and powerful, thus enabling 'deterrence by punishment'. Another condition is that, unlike the traditional notion of deterrence, cyber-deterrence cannot be absolute. This means that some kinds of actor and some types of action cannot be deterred. Deterrence is most likely to work when attempting to deter cyberattacks that would have severe consequences and strategic purposes. Normally such attacks are planned and conducted by states.

A related condition is that the problem of attribution of an attack can be resolved to a degree when the strategic context and operational reality are taken into consideration.[3] This leads to greater confidence that a cyberattack has been initiated by a state actor, as was the case with the use of the Stuxnet worm against Iranian nuclear facilities.[4] The level of intelligence and technological capabilities required to carry out such an attack narrows down the list of suspects. In other words, attribution becomes less of a problem if the target of deterrence is a state, particularly a powerful one, and the behaviour to be deterred is a sophisticated, strategic cyberattack against an air-gapped facility. This has implications throughout civilian and military nuclear infrastructure.

Considering these conditions, the impact of machine learning on cyber-deterrence is ambiguous. On the one hand, cyber-deterrence may be enhanced. A typical example is in cyber-defence. By providing more active and adaptive defence, reducing human effort in monitoring threats and generating a timelier response, machine learning may raise the cost for a potential attacker and thus help promote deterrence by denial in cyberspace. Attribution is another area that machine learning may bolster. Deterrence would seem more credible if the attacker were to lose its anonymity. As such, cyber-deterrence could be more feasible with the intervention of machine learning.

[2] Brantley (note 1); Haggman, A., 'Cyber deterrence theory and practise', eds M. Lehto and P. Neittaanmäki, *Cyber Security: Power and Technology* (Springer: Cham, 2018), pp. 63–81; and Bendiek, A. and Metzger, T., 'Deterrence theory in the cyber-century: lessons from a state-of-the-art literature review', eds D. W. Cunningham et al., *Informatik 2015: Informatik, Energie und Umwelt* [Informatics 2015: computer science, energy and environment] (Gesellschaft für Informatik: Bonn, 2015), pp. 553–70.

[3] Lindsay, J. R., 'Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack', *Journal of Cybersecurity*, vol. 1, no. 1 (2015), pp. 53–67.

[4] 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.

III. Problems for cyber-deterrence

Even if the above conditions are met, several factors may also reduce the effectiveness of cyber-deterrence.

The first is the possibility of adversarial machine learning. As cyber-defence models based on machine learning become more effective at detecting threats, potential attackers may look for ways to confuse the models. This is often called adversarial machine learning. Even if actors on the defensive side can rely on AI models to safeguard their systems, their confidence in deterrence by denial must not be exaggerated. This is because offenders may succeed in poisoning the models (also known as machine learning poisoning) or may find other ways to evade them. In this sense, the security benefits offered by AI could be offset. However, a 2018 report warned that relatively little attention has been paid to making AI-based defences robust against attackers that anticipate their use.[5] Ironically, the use of machine learning for cyber-defence can actually expand the attack surface of a defence system—the points at which an attacker can interact with the system—due to this lack of attention and other vulnerabilities.

A second problem is the blurred connection between actors and capabilities. Strategic cyberattacks—attacks that inflict damage of strategic national impact—are currently unlikely to be conducted by individuals. Cyber operations that intend to change the target's behaviour or to make the target bear considerable losses often involve complex efforts for preparation, organization, coordination, and testing and rehearsal. They necessitate an abundance of critical resources, such as discovery of zero-day vulnerabilities, hacking tools and talent. Such cyber operations also require adequate information about the target systems' defence preparedness. However, with the development of machine learning, these efforts could be performed or facilitated by automated and adaptive programmes. These programmes would be able to dig up vulnerabilities, circumvent detection, defeat anti-malware systems or even redesign an operation according to the recognized properties of the target system. This means that complex cyber operations may not require the same level of organizational complexity in the AI era. Individual hackers or small groups could also complete tasks that have strategic consequences. This would create several knock-on problems for cyber-deterrence. First, the 'known identity plus known demand' condition would be more difficult to establish—that is, being able to determine both the source of the attack and its aims would be muddied.[6] Second, attribution would become more difficult because capabilities, including other forensic evidence such as language or similarity with past operations, may no longer be a reliable indicator for attribution. Third, cyberattacks with severe consequences may proliferate, undermining strategic stability in cyberspace.

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

[6] Borghard, E. D. and Lonergan, S. W., 'The logic of coercion in cyberspace', *Security Studies*, vol. 26, no. 3 (2017), pp. 452–81.

Finally, increased use of AI in cybersecurity would make the gathering and sharing of data even more important. This may make alliance politics in cyberspace more common and intensive. Machine learning-based cyber-defence normally takes two forms. One is supervised learning, where the goal is to learn from known threats and to generalize and apply this knowledge to new threats. The other is unsupervised learning, in which programmes try to find suspicious deviations from normal behaviour. Either form would require extensive analysis of data and strong intelligence capabilities and networks. Therefore, to make deterrence effective, a state would need to cooperate with other states in information sharing in order to build a global intelligence network. This may encourage alliance relationships in cyberspace. The negative outcome would be an intensification of the already evident cleavages in cyberspace, creating a sense of antagonism between different groups and making global consensus on cybersecurity norms even harder to reach.

IV. Conclusions

Overall, AI and machine learning pose risks and offer benefits to cybersecurity. The impact on cyber-deterrence remains unclear, since both defence and offence could be buttressed by the development of AI. This suggests the need for more dialogue among AI researchers, strategic researchers, policymakers and other relevant stakeholders to reach greater clarity on cyber-deterrence and how it may have an impact on future strategic relations and arsenals.

4. Integration of neural networks into hypersonic glide vehicles

LORA SAALMAN*

Chinese researchers confront many of the same hurdles faced by other contenders aiming to integrate neural networks into such kinetic platforms as hypersonic glide vehicles.[1] Pre-existing foreign research on applications of neural networks in missile seekers, missile fusing, sonar target discrimination, automatic target recognition and auto piloting has argued that neural networks are 'high risk–high payoff'.[2] China appears willing to take on this challenge involving artificial intelligence (AI). Over the past decade, its experts have advanced beyond foreign publications to a prolific release of domestic studies that pursue the benefits of integrating neural networks to enhance manoeuvrability and to penetrate defences, thereby reshaping conventional and nuclear deterrence dynamics.[3]

This essay provides a brief overview of Chinese technological writings on neural networks and hypersonic glide vehicles. It begins (in section I) with the foundations of cross-collaboration and innovation in Chinese research institutes on these AI-enabled technologies. The essay continues (in section II) by detailing how these developments in neural network integration and hypersonic vehicles are reshaping conventional and nuclear deterrence dynamics. It concludes (in section III) with an analysis of how these technological advances may also mark a shift towards a more offence-oriented version of 'active defence' (积极防御) in the future.

I. Expanding collaboration and innovation

Hypersonic glide vehicles are a growing factor in strategic stability calculations, which first gained popularity in the debates over the United States' Conventional Prompt Global Strike (CPGS) programme.[4] These systems are characterized by speed, precision and manoeuvrability that can be applied to defeat missile defences and to deliver either a conventional or a nuclear payload. On reaching near space, the vehicle is ejected from its boosters to begin its glide phase, during

[1] 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*, Strategic Multilayer Assessment (SMA) Periodic Publication (Department of Defense: Washington, DC, Dec. 2018), pp. 153–60.

[2] Webster, W. P., 'Artificial neural networks and their application to weapons', *Naval Engineers Journal*, vol. 103, no. 3 (May 1991), pp. 46–59.

[3] Ma, G. et al., 'Adaptive backstepping-based neural network control for hypersonic reentry vehicle with input constraints', *IEEE Access*, vol. 6 (2018), pp. 1954–66.

[4] Saalman, L., 'Prompt global strike: China and the spear', Independent faculty article, Asia–Pacific Center for Security Studies, Apr. 2014. On CPGS see also chapters 13 and 14 in this volume.

* The views expressed in this essay are those of the author and do not necessarily reflect those of any organization to which she is affiliated.

which it can accelerate to more than Mach 5 (6125 kilometres per hour). The glide phase allows it to manoeuvre aerodynamically to evade interception and extends its range. Unlike conventional re-entry vehicles, which follow a predictable ballistic trajectory, hypersonic glide vehicles are almost impossible to intercept using conventional missile defence tracking systems.[5] In Chinese writings, the definition of these vehicles is broad and includes a wide variety of near-space and other platforms.[6]

Based on a review of over 300 articles from universities and military institutes in China that builds on a database of 3000 Chinese-language papers on hypersonic and AI advances, it can be concluded that domestic and international cross-institute collaboration on these technologies has become increasingly common.[7] Engineers from the People's Liberation Army (PLA) Rocket Force, the College of Mechatronic Engineering and Automation of the National University of Defense Technology, Harbin University, Tsinghua University, Beihang University, the China Academy of Launch Vehicle Technology, the PLA Rocket Force Engineering University, Northwestern Polytechnical University and the Beijing Institute of Tracking and Telecommunications Technology, among other institutions, have been working—often collectively—to find new solutions for some of the more intractable problems faced in control dynamics of hypersonic glide vehicles.[8]

One example of a joint project at the domestic level is a paper on neural networks and hypersonic flight dynamics by four authors from the geographically diverse Naval Aeronautical and Astronautical University in Yantai, Zhejiang University in Hangzhou and China State Shipbuilding Corporation in Beijing.[9] An example at the international level is a study on adaptive control for near-space hypersonic vehicles authored by three Chinese researchers based at Nanjing University of Aeronautics and Astronautics and a Chinese colleague at the University

[5] Saalman, L., 'Factoring Russia into the US–Chinese equation on hypersonic glide vehicles', SIPRI Insights on Peace and Security no. 2017/1, Jan. 2017.

[6] Saalman (note 4).

[7] Liu, Q. (刘清楷) et al., '高超声速飞行器俯冲段制导控制方法研究' [Guidance and control design for hypersonic vehicle in dive phase], 现代防御技术 [Modern Defence Technology], vol. 45, no. 6 (Dec. 2017), pp. 74–81; Zhang, J. et al., 'Adaptive sliding mode control for re-entry attitude of near space hypersonic vehicle based on backstepping design', *IEEE/CAA Journal of Automatica Sinica*, vol. 2, no. 1 (Jan. 2015), pp. 94–101; Bu, X. (卜祥伟) and Wang, K. (王柯), '高超声速飞行器输入受限自适应反演控制研究' [Study on adaptive backstepping control of hypersonic vehicles with input constraints], 上海航天 [Aerospace Shanghai], vol. 34, no. 6 (June 2017), pp. 26–35; and Ma, Y. (马宇) and Cai, Y. (蔡远利), '面向高超声速飞行器的新型复合神经网络预测控制方法' [A novel composite model predictive control method based on neural network for hypersonic vehicles], 西安交通大学学报 [Journal of Xi'an Jiaotong University], vol. 51, no. 6 (June 2017), pp. 28–34.

[8] Saalman (note 4); Pan, L. (潘亮) et al., '高超声速飞行器滑翔制导方法综述' [A survey of gliding guidance methods for hypersonic vehicles],], 国防科技大学学 [Journal of National University of Defense Technology], vol. 39. no. 3 (June 2017), pp. 15–22; Zhang, K. (张凯) and Xiong, J. (熊家军), '高超声速滑翔目标多层递阶轨迹预测' [Multi-level recursive trajectory prediction for hypersonic gliding reentry vehicle], 现代防御技术 [Modern Defence Technology], vol. 46, no. 4 (Aug. 2018), pp. 92–98; Wang, M. (王明昊), Liu, G. (刘刚) and Hou, H. (侯洪庆), '吸气式高超声速飞行器纵向通道控制器设计研究' [Design of longitudinal controller for air-breathing hypersonic vehicle], 计算机测量与控制 [Computer Measurement & Control], vol. 21, no. 4 (Apr. 2013), pp. 955–58; Yao, C. (姚从潮) et al., '一种高超音速飞行器轨迹线性化控制方法研究' [Research on trajectory linearization control method for hypersonic vehicles], 计算机仿真 [Computer Simulation], vol. 29, no. 12 (Dec. 2012), pp. 80–85.

[9] Wang, S. et al., 'Neural control of hypersonic flight dynamics with actuator fault and constraints', *Science China Information Sciences*, vol. 58, no. 7 (July 2015).

of Virginia, USA.[10] Not only does their work detail China's integration of neural network-based control into their hypersonic programmes, but it also reveals information on support systems and facilities for testing.

These works constitute a sizeable advance in Chinese transparency and collaboration, which are crucial for technological advancement. Beyond the use of models traditionally used in other countries—such as Lyapunov stability theory or the Singer model[11]—Chinese researchers are now developing their own models for robust non-linear adaptive control systems that integrate terminal sliding mode controls, predictive controls, fuzzy neural network controls and non-linear dynamic inverse controls.[12] These AI-based controls are meant to address the hypersonic glide vehicle's high flight envelope, complex flight environment, severe non-linearity, intense and rapid time-variance, dynamic uncertainty during the dive phase and strong coupling characteristics.

Given the need for greater resilience in the absence of data, a number of Chinese articles seek to integrate 'radial basis function neural networks' (基于径向基函数的神经网络) to mitigate non-linearity and uncertainty in physical and aerodynamic parameters.[13] Further, Chinese experts are also applying bee colony algorithms and swarm technology to address parameter-identification problems found in complex operating environments.[14] These works rely on autonomy as a means of achieving coordinated guidance control of space-based hypersonic vehicles in proximity, namely 'cooperative guidance and control of hypersonic vehicle autonomous formation' (高超声速飞行器自主编队协同制导控制).[15]

Thus, whether integrated into autonomous swarm-like formations or used for enhancement of manoeuvrability and control, neural networks are contributing to China's communication and decision-support systems, high-precision guidance,

[10] Zhen, Z. (甄子洋) et al., '基于自适应控制的近空间高超声速飞行器研究进展' [Research progress of adaptive control for hypersonic vehicles in near space], 宇航学报 [Journal of Astronautics], vol. 39, no. 4 (Apr. 2018), pp. 355–67.

[11] Zhang et al. (note 7); and Wei, X. (魏喜庆) et al., '基于Singer模型的高超声速飞行器轨迹跟踪与预测' [Hypersonic vehicle trajectory tracking and prediction based on the Singer model], 航天控制 [Aerospace Control], vol. 35, no. 4 (Apr. 2017), pp. 57–61.

[12] Wang et al. (note 8); Guo, X.(郭相科) et al., '一种新的临空高超声速飞行器滑跃段跟踪算法' [A new tracking algorithm for near space hypersonic vehicle in gliding jumping phase], 宇航学报 [Journal of Astronautics], vol. 38, no. 9 (Sep. 2017), pp. 971–78; and Hu, C. (胡超芳) et al., '高超声速飞行器模糊自适应动态面容错控制' [Fuzzy adaptive dynamic surface fault-tolerant control for hypersonic vehicles], 天津大学学报 (自然科学与工程技术版) [Journal of Tianjin University (Science and Technology)], vol. 50, no. 5 (May 2017), pp. 491–95.

[13] Ma and Cai (note 7); Wang, F. (王芳), '基于反步法的高超声速飞行器鲁棒自适应控制' [Robust adaptive control of hypersonic vehicle based on backstep method], Dissertation, Tianjin University, 2014; and Yao et al. (note 8).

[14] Li, S. (李霜天) and Duan, H. (段海滨), '基于人工蜂群优化的高超声速飞行器在线参数辨识' [Artificial bee colony approach to online parameters identification for hypersonic vehicle], 中国科学: 信息科学 [Scientia Sinica Informationis], vol. 42, no. 11 (Nov. 2012), pp. 1350–63.

[15] Fan, C. (樊晨霄) et al., '临近空间高超声速飞行器协同制导控制总体技术研究' [System design of cooperative guidance and control of near space hypersonic vehicles], 战术导弹技术 [Tactical Missile Technology], Apr. 2018, pp. 52–58; and Zong, Q. (宗群), '高超声速飞行器建模与自主控制技术研究进展' [New development of modeling and autonomous control for hypersonic vehicle], 科技导报 [Science & Technology Review], vol. 35, no. 21 (May 2017), pp. 95–106.

targeting and discrimination, as well as cyber-centric and electronic warfare.[16] Understanding this range of capabilities is crucial since they have the potential to be game changing when applied in either conventional or nuclear deterrence.

II. Reshaping deterrence with neural networks and hypersonic glide

The advances detailed above have vast implications beyond simply enhancements to existing capabilities in manoeuvrability and targeting. They are reshaping strategic dynamics. In a 2018 article on near-space hypersonic vehicles, five Chinese researchers working at the Beijing 'Long March Vehicle' Institute of Space, the Wanyuan Science and Technology Company and the School of Aeronautics and Astronautics of Sun Yat-sen University discuss the strategic rationale behind these technological developments. They describe the importance of these vehicles to counter 'the rapid development of modern land, sea, air and space integrated defence technologies, in particular the regional air defence systems and short-range defence forces of high-value military targets (such as aircraft carrier battle groups and strategic command centres) that constitute a multilayered anti-missile air defence system'.[17] In the case of the USA, such a multilayered system represents a confluence of its deployments and alliance structures in East Asia and its national deterrence aims.

As such, Chinese hypersonic glide platforms should not be thought of as simply short-range or intermediate-range platforms.[18] Over a third of the recent Chinese analyses and research surveyed focuses on 'near-space' (临近) hypersonic vehicles and some even explicitly focus on 'global approximation' (全局逼近) capabilities for non-linear hypersonic vehicle systems.[19] While Chinese analyses give the US X-43A and X-51 unmanned hypersonic vehicles as examples of these wide-ranging platforms, China's technical journals are pursuing the same sets of technologies to be able to conduct stealthy, rapid and overwhelming strikes against advanced defence systems and high-value military targets.[20]

Beyond the physical range that near-space flight allows, these studies also focus on the pursuit of flexible and large-scale operations, with a stated goal of improving the effectiveness of guided-weapon systems. They seek to develop relatively low-cost guided weapons and equipment by applying neural network-enabled guidance to form synergistic electronic countermeasures, cascade joint

[16] Zhao, H. (赵贺伟) et al., '高超声速飞行器自适应神经网络控制' [Adaptive neural network controller design for hypersonic vehicle], 固体火箭技术 [Journal of Solid Rocket Technology], vol. 40, no. 2 (Feb. 2017), pp. 257–63.

[17] Fan et al. (note 15), p. 53 (author translation). See also Ren, Z. (任章) and Yu, J. (于江龙), '多临近空间拦截器编队拦截自主协同制导控制技术研究' [Research on the autonomous cooperative guidance control for the formation interception of multiple near space interceptors], 导航定位与授时 [Navigation Positioning & Timing], vol. 5, no. 2 (Mar. 2018), pp. 1–6.

[18] Wang, Q. (王庆洋) et al., '临近空间高超声速飞行器气动力及气动热研究现状' [Research status on aero-dynamic force and heat of near space hypersonic flight vehicles], 气体物 [Physics of Gases], vol. 2, no. 4 (July 2017).

[19] Zhen et al. (note 10).

[20] Fan et al. (note 15); and Zhen et al. (note 10).

penetration and scalable saturation strikes.[21] These studies are not simply about anti-access/area-denial (A2/AD) and regional dynamics: their longer-term aim is evasion and penetration of US defences, both regionally and globally.

III. Conclusions

Whereas a decade ago Chinese technical articles focused on developing 'countermeasures' (对策) against hypersonic glide vehicles, the majority of recent papers written in China are preoccupied with research and development of AI-enabled offensive platforms. In fact, among the hundreds of Chinese-language articles and papers surveyed for this essay, only one has an explicit focus on enhancing hypersonic glide intercept.[22] The vast majority of technical analyses and designs instead seek to penetrate missile defences. After spending a decade on countering US plans for CPGS, it is not surprising to witness this Chinese shift towards offence in domestic research and priorities.

Rather than bolstering China's concept of 'active defence', this predominance of offensive platforms suggests a trend towards a more offence-oriented stance. China has long hedged when it comes to the payload of its hypersonic glide systems, placing it somewhere between Russia's emphasis on nuclear warheads and the USA's focus on conventional warheads.[23] Yet the very aim of defeating US missile defences suggests that China's hypersonic vehicles have a strong potential to be used for a nuclear payload in the future. The present author's recent interactions with PLA generals and admirals has re-emphasized this point when it comes to the evolution of China's own hypersonic glide vehicles.

In sum, it cannot be taken for granted that China's neural network and hypersonic glide vehicle developments are only conventional. With the diminishing number of Chinese technical journal papers that seek countermeasures and the increasing number that explore deployment of near-space neural network-enabled hypersonic glide platforms, China's tactical and strategic orientation is shifting towards an offensive one, whether or not this is explicitly stated in Chinese rhetoric or official military doctrine. This marks a direct confluence of not simply China's hypersonic vehicles and neural networks, but also its concepts of conventional and nuclear deterrence.

[21] Fan et al. (note 15); and Zhen et al. (note 10).
[22] Ren and Yu (note 17).
[23] Saalman (note 5).

5. Applications of machine learning in North Korea and South Korea

HWANG JI-HWAN*

The application of artificial intelligence (AI) in nuclear and military arenas is extremely confidential in both the Democratic People's Republic of Korea (DPRK, or North Korea) and the Republic of Korea (South Korea). However, some information is available in open sources. This essay gives a brief overview of the open source data on advances in machine learning for each of the two countries, with a particular focus on machine learning. It begins (in section I) with a discussion of various AI-enabled platforms under development by South Korea. It then turns (in section II) to North Korean machine learning developments by reviewing the available open source materials.

I. South Korea

The South Korean Government has recently been actively involved with what is known as the fourth industrial revolution—characterized as the merger of physical, biological and cyber technologies.[1] As part of this drive, its Ministry of Science and Information and Communications Technology has outlined what could be called South Korea's digital 'ICBM', encompassing the Internet of things, Cloud computing, Big data and Mobile technology.[2] This work proceeds under the auspices of the Intelligence Information Task Force. This grouping has also undertaken cooperation with the Ministry of National Defense (MND) on the application of these new technologies in military affairs, with a particular focus on integrating AI into weaponry and defence management.

Nonetheless, South Korea is in the initial stages of the military application of machine learning and continues to lag behind other countries. In large part, this is due to the lack of AI-related investment and research and development. It is further hampered by the paucity of big data, which is instrumental to advancing machine learning applications within the military. Given the relative newness of these technologies, the South Korean Government and companies have not had enough time to accumulate the requisite stores of information or big data, much less to develop the necessary domestic laws and institutions to govern these advances. There is also a serious lack of military–industrial cooperation and a

[1] See e.g. South Korean Ministry of Science and Information and Communications Technology, 'Policies'. On the 4th industrial revolution see e.g. Schwab, K., The Fourth Industrial Revolution (Penguin: London, 2017).

[2] South Korean Ministry of Science and ICT, 'The Republic of Korea rides on the big road of the 4th industrial revolution', accessed 10 Aug. 2019.

* The views expressed in this essay are those of the author and do not necessarily reflect those of any organization to which he is affiliated.

lack of machine learning expertise in the non-governmental areas in which most machine learning research is being conducted. Finally, South Korea faces obstacles due to slow bureaucratic procedures that delay the development of machine learning and other AI applications in both governmental and non-governmental arenas, sometimes resulting in a rejection of these new technologies.

As a result of these impediments, South Korea lacks a comprehensive military machine learning project such as the United States' Project Maven, which aims to develop and integrate 'computer-vision algorithms needed to help military and civilian analysts encumbered by the sheer volume of full-motion video data that [the US Department of Defense] collects every day in support of counter-insurgency and counterterrorism operations'.[3] While the South Korean Government and companies are interested in this project, they lack the capacity to develop a similar one. Nonetheless, the MND does have some smaller-scale machine learning and more general AI-related projects. For example, it is developing an unmanned combat vehicle, known as the Gyun-Ma (견마) robot, for reported use in detection, communications, surveillance and reconnaissance.[4]

South Korea has also been working on two projects on AI software that engages in information processing and response: Exobrain, a government-funded project, and ADAMs, a private initiative by Saltlux, a cognitive computing company.[5] These illustrate the role of both government and private industry, without offering much in the way of detail on collaboration. Nonetheless, ADAMs possesses the intelligence for linguistic, visual, emotional and reasoning capability that may be applied in command, control, communications, computers, intelligence, surveillance and reconnaissance (C4ISR). South Korea has also conducted research into omni-directional movement interactive software technology to support virtual soldier exercises.[6] This initiative is intended to build a simulated combat environment for soldiers by employing big data-supplied machine learning technology.

All of the above-mentioned technologies are still in their initial stages and face many obstacles to their employment in the South Korean military. However, given South Korea's strong background in information technology, it is well-positioned to build advanced military-related machine learning and other types of AI technology in the near future.

[3] Shanahan, J., US Department of Defense, 'Disruption in UAS: the Algorithmic Warfare Cross-Functional Team (Project Maven)', Presentation, Royal Australian Air Force, Airpower Development Centre, 20 Mar. 2018; and Pellerin, C., 'Project Maven industry day pursues artificial intelligence for DOD challenges', US Department of Defense, 27 Oct. 2017.

[4] Kim, M., 'The South Korean military is catching up with a new drone army', *Security Times*, Feb. 2018; and Shin, J., 'S. Korea aims to become no. 4 robotics player by 2023', *Korea Herald*, 22 Mar. 2019.

[5] Prakash, A., 'South Korean AI sees continued development, investments', Robotics Business Review, 1 May 2018; and Park, S., 'Saltlux unveils artificial intelligence software ADAMs: Yonhap', *Aju Business Daily*, 23 Nov. 2016.

[6] Global Sources, 'Products from ED Co., Ltd: omni-directional mobile robot'.

II. North Korea

Much as in other areas, North Korea tends to be a black box when it comes to development of AI and applications of machine learning. However, it is also reported to maintain a strong interest in these technologies. Even if it has made efforts to apply this new technology in military affairs, this information is highly confidential given the security environment on the Korean peninsula. Nonetheless, the North Korean Government has recently emphasized the importance of information and communications technology (ICT). In particular, the North Korean leader, Kim Jong Un, emphasizes advanced ICT as one of the main pillars of the country's economic development, meriting priority in its development efforts.[7] Furthermore, given North Korea's substantial efforts in cyberspace, it is likely to have sought to develop machine learning and AI more broadly for military uses. Thus, while these technologies appear to be in the early stages of development, they have started to become evident in non-military arenas, with dual-use military implications.

As one example, North Korea has started to develop its own AI technology-based version of the game go and has sought to apply the technology to other areas.[8] North Korea is also thought to have held ICT contests to encourage expansion of this skill set among its youth. The College of Computer Science of Kim Il Sung University has also developed its own speech-recognition system, Ryongnamsan (룡남산) 5.1, and the Natural Science Institute has developed a fingerprint and facial recognition system.[9] Kim Chaek University of Technology has developed a multilingual interpretation programme, known as Genius (신동). The AI Institute of the Korea Computer Center is also reportedly leading AI and related machine learning developments in North Korea.[10]

While notably lacking in transparency, some research on machine learning and AI has been published in the scientific sections of the *Kim Il Sung University Journal*. In particular, this journal recently published an article on audio classification using a deep belief network (DBN), an example of machine learning.[11] This demonstrates the existence and level of the North Korean research. It also shows that North Korean machine learning technology is at an initial stage since it imitates foreign research. Much remains based on the earlier research of the Canadian academic Geoffrey Hinton on fast learning algorithms for

[7] Jakhar, P., 'North Korea's high-tech pursuits: propaganda or progress?', BBC, 15 Dec. 2018.

[8] Choi, S., 'North Korea's artificial intelligence go software', Korea IT Times, 4 Jan. 2011.

[9] Ji, D., 'Facial, voice recognition software on display at North Korean IT exhibit', NK News.org, 23 Nov. 2017.

[10] Kang, J. (강진규), '김정은 시대 북한 IT 현황과 기술 수준' [North Korea's IT status and technology level under Kim Jong Un], Digital Hurricane, 17 May 2018.

[11] Ri, J. (리정철) and Hyon, S. (현성군), '음소음성인식에서 심층신뢰망을 리용한 한가지 음향모형화 방법' [An acoustic modeling method based on Deep Belief Networks in the phone speech recognition], 김일성종합대학학보: 자연과학 [Kim Il Sung University Journal: Natural Science], vol. 62, no. 8 (Aug. 2016), pp. 30–34.

DBNs.[12] Nonetheless, acquisition and integration of machine learning requires a foundation. Thus, while available sources suggest that technology in North Korea is well behind that of South Korea, its rapid advances in cyber operations and ICT indicate that it can be expected in the near future to develop machine learning and other types of AI technology and to apply them in military affairs.

[12] On Hinton see Boulanin, V., 'Artificial intelligence: a primer', ed. V. Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. I, *Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019), pp. 13–25, p. 17.

6. Military developments in artificial intelligence and their impact on the Korean peninsula

SU FEI*

Applications of artificial intelligence (AI) in the military domain by the Democratic People's Republic of Korea (DPRK, or North Korea) and the Republic of Korea (South Korea) focus on two platforms: unmanned systems and cyber technologies. Given that one of the major security concerns on the Korean peninsula remains North Korea's nuclear programme, it is crucial to understand how AI applications in these two platforms may affect nuclear risk and strategic stability on the Korean peninsula.

This essay continues (in section I) by detailing South Korea's AI-enabled advances in unmanned systems and cyberspace. It then (in section II) reviews North Korea's advances in artificial neural networks, robotics, unmanned systems and cyberspace based on a review of journal articles. Using this technological foundation, the essay then analyses (in section III) how these developments may have an impact on stability on the Korean peninsula at the nuclear level. It concludes (in section IV) by discussing the short- and longer-term prognoses if proliferation of dual-use AI technology occurs.

I. South Korea

Efforts to apply AI for defensive use can be seen in both the South Korean military and cooperation between the public and private sectors.

At the beginning of 2019, the South Korean Army launched an Artificial Intelligence Research and Development Center.[1] The short-term aim of this centre is to build the vision and concept for military applications of AI and to develop the next generation of combat power.

To the same end, the Korea Advanced Institute of Science and Technology (KAIST) announced in February 2018 that it would team up with Hanwha Systems, the defence business unit of Hanwha Group, to research and develop AI weapons.[2] Their cooperation on military AI technology includes four projects: navigation algorithms for large-scale unmanned underwater vehicles (UUVs), AI-based command systems, AI-based aviation training systems and AI-based object-tracking techniques. To support these endeavours, KAIST established the Research Center for the Convergence of National Defense and Artificial Intelligence with the support of Hanwha Systems. In addition to the above four

[1] Dominguez, G., 'RoKAF to launch AI research centre', *Jane's Defence Weekly*, 31 Dec. 2018.

[2] Hanwha Systems, '한화시스템–KAIST, 국방 AI기술 개발 나선다' [Hanwha Systems–KAIST step up the development of defence AI technology], 20 Feb. 2018.

* The views expressed in this essay are those of the author and do not necessarily reflect those of any organization to which she is affiliated.

projects, this centre also conducts research related to cyberwarfare and AI robotics.[3]

Underlying this, the South Korean Ministry of National Defense (MND) has identified unmanned aerial vehicles (UAVs), including high-altitude UAVs and unmanned combat aerial vehicles (UCAVs), as one of the new domains for strengthening combat capability.[4] The South Korean Army also announced in 2017 its intention to establish a specialized unit, named Dronebot Jeontudan (드론봇 전투단), to operate UAVs and unmanned ground vehicles.[5]

South Korea has at least three core reasons for seeking to adopt unmanned systems into its military. First, the reach of these platforms into complex terrain such as the high and steep mountains of the Korean peninsula can provide better surveillance and reduce miscalculation when conflict occurs. Second, South Korea's neighbours, including North Korea, have already been operating unmanned systems and swarms. In 2014 and 2016, North Korean UAVs were already spotted intruding into South Korean airspace.[6] Third, UAVs can be used to counter North Korea's non-nuclear provocations without human casualties.[7]

Another domain to which South Korea has paid particular focus is cyberwarfare. This is in large part due to the frequent cyberattacks that the country has faced, which have been linked to North Korea both through official government channels and via technology and software companies.[8] A recent example occurred in January 2019, when a group of reporters covering the South Korean Ministry of Unification, which is in charge of relations with North Korea, received a malware-ridden invitation to a press conference for the second summit between President Donald J. Trump of the United States and the North Korean leader, Kim Jong Un.[9] While the ministry is still investigating the incident, a Seoul-based software development company, ESTsecurity, has already attributed the hack to North Korean sources.

To prevent and counter cyberattacks, South Korea is considering development of AI-based means for early detection and response to cyberthreats.[10] South Korea's first National Cybersecurity Strategy, issued in April 2019, explicitly states that it will 'Expand the scope of detecting cyberattacks to enable real-time detection

[3] Korea Advanced Institute of Science and Technology (KAIST), 'KAIST 2018 안보/국방 융합4.0포럼 환영사' [Welcome to the KAIST 2018 Security—Defence Convergence 4.0 Forum], 23 Mar. 2018.

[4] South Korean Ministry of National Defense (MND), *2019*년 국방부 업무보고: 국민과 함께—평화를 만드는 강한 국방 [2019 Ministry of National Defense report: with the people—strong defence to make peace] (MND: Seoul, 20 Dec. 2018).

[5] Wong, K., 'DX Korea 2018: RoKA outlines plans for new "Dronebot Warrior" unit', *Jane's Defence Weekly*, 14 Sep. 2018.

[6] Ahn, J. H., 'North Korean drone spotted near border', NK News.org, 13 Jan. 2016.

[7] Hironaka, M. and Yoon, S., 'Proliferated drones: a perspective on Japan', Center for New American Security, [June 2016].

[8] Park P., 'Experts examine Asia's approach to cybersecurity', Order from Chaos, Brookings, 28 Aug. 2018.

[9] 'North Korea-backed hackers intensify information warfare, financial theft', *Korea Herald*, 26 Mar. 2019.

[10] South Korean Ministry of National Defense (note 4).

and blocking and develop AI-based response technologies'.[11] In particular, the Korea Internet and Security Agency (KISA) has been using big data and deep learning technologies to better understand how machines can identify and detect threats based on collected information. It has also focused on the use of autonomous systems to analyse vulnerabilities and to generate possible solutions for the attacks. Finally, it has worked on predicting cyberthreats based on known information.[12]

II. North Korea

North Korea's interest in developing AI technology is already evident in civilian fields and there is great potential for this technology to be integrated into weapon systems. One of the most probable applications is through enhanced cyber capabilities. An academic paper published in 2018 in the *Kim Il Sung University Journal* demonstrates that North Korea is researching improvements in detection of intrusive cyber operations via artificial neural networks and genetic algorithms.[13] In addition to defensive capabilities, there are also concerns that North Korea will develop the ability to conduct AI-enabled cyberattacks.[14]

North Korea has proved that it is capable of and willing to use its cyber capabilities to conduct disruptive cyber operations. It has allegedly been involved in a number of cyber incidents, such as the 2014 cyberattack on Sony Pictures and the 2017 WannaCry incident that affected over 150 countries.[15] When North Korea is faced with international sanctions, its cyber operations have brought financial gain and inserted this small and isolated country further into international political discourse. Cyber operations represent a cost-effective means and opportunity for North Korea to project its power, despite the limited size and strength of its military. It is critical to North Korea's national strategy and there be greater investment in this field in the years to come.[16]

Outside the cyber domain, North Korea is also focusing on developments in robotics. A research report issued by Kim Il Sung University in 2018 reveals that

[11] South Korean National Security Office, *National Cybersecurity Strategy* (National Security Office: Seoul, Apr. 2019, p. 16.

[12] Korea Internet and Security Agency (KISA), '정보보호 원천기술 개발 및 보급' [Development and dissemination of information safety technology], accessed 23 May 2019.

[13] Pak, S. (박성호) and Hwang, C. (황철진), '망침입검출에서 속성선택에 의한 성능개선' [Performance improvement by attribute selection in the network intrusion detection system], 김일성종합대학학보: 정보과학 [Kim Il Sung University Journal: Information Science], vol. 64, no. 2 (2018), pp. 34–39. See also Kang, J. (강진규) '북한, 보안에 AI 적용을 추진하고 있다' [North Korea is pushing AI into security], NK경제 [NK Economy], 6 Nov. 2018.

[14] '中露、サイバー攻撃にＡＩ活用　北も能力獲得か　手口を学習、標的選定も　元在日米軍司令部サイバーセキュリティー長が証言' [China and Russia use AI for cyberattacks—North also learns how to acquire ability, learns tricks and selects targets], Sankei Shimbun, 14 Feb. 2018; and Goud N., 'North Korea, China, and Russia to launch hyper war says NATO', Cybersecurity Insiders, accessed 25 July 2019.

[15] US Department of Justice, 'North Korean regime-backed programmer charged with conspiracy to conduct multiple cyber-attacks and intrusions', 6 Sep. 2018.

[16] Jun, J., LaFoy, S. and Sohn, E., *North Korea's Cyber Operations: Strategy and Reponses* (Center for Strategic and International Studies: Washington, DC, Dec. 2015); and Ko, L., 'North Korea as a geopolitical and cyber actor', New America, 6 June 2018.

North Korea is working on the application of neural networks in autonomous mobile robots.[17] Another paper shows that North Korea is studying how to measure distance and recognize obstacles when operating autonomous robots.[18] Such technologies can be potentially used for improving its UAV capabilities. For example, the application of neural networks by using imagery databases can enable better assessment of the surrounding environment.[19]

In fact, North Korea's interests in developing UAVs can be traced back to the 1970s. Over the past decades it has made steady efforts to produce and enhance its UAV capability by obtaining foreign technology and importing foreign products.[20] It is unclear how many unmanned platforms North Korea possesses in total. A 2016 report for the United Nations Security Council estimates the number at 300, while in 2017 South Korean experts estimated 1000.[21] Further, self-destructing UAVs were displayed at a military parade in 2013.[22]

North Korea has also been working on miniaturization of UAVs. In 2016 it reduced the size to one metre, which is about one-third the size of UAVs that crashed in South Korea in 2014.[23] Such small platforms are difficult for South Korean radars to detect. However, with South Korea's efforts to deploy counter-UAV systems to detect these small devices, the extent to which this autonomous technology may affect reconnaissance remains to be seen.[24] One of the obstacles faced by North Korea's UAVs, at least in the case of the one found in 2017 in South Korea, is the lack of an autonomous landing guidance system.[25]

[17] Kang, J. (강진규), '북한, AI 적용 이동형 로봇 연구 중 [North Korea is studying applying AI technology in mobile robots]', NK경제 [NK Economy], 4 Oct. 2018.

[18] Han, H. (한학수) and Choe M. (최명성), '안내로보트의 항행을 위한 촬영기와 레이자 거리수감부의 교정에 대한 연구' [Research of extrinsic calibration of a camera and a 2D laser range sensor for navigation of guided robot], 김일성종합대학학보: 자연과학 [Kim Il Sung University Journal: Natural Science], vol. 63, no. 12 (Dec. 2016), pp. 39–41. See also Kang T., 'North Korea's quest for autonomous technology', *The Diplomat*, 13 July 2018.

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

[20] Bermudez, J. S., 'North Korea drones on', 38 North, 1 July 2014.

[21] United Nation, Security Council, Report of the Panel of Experts submitted pursuant to Resolution 1874 (2015), 18 Jan. 2016, S/2016/157, 24 Feb. 2016; and Chung, G. (정구연) and Lee, K. (이기태), 과학기술발전과 북한의 새로운 위협: 사이버 위협과 무인기 침투 [Science and technology development and new threats in North Korea: cyberthreats and unmanned infiltration] (Korean Institute for National Unification: Seoul, Dec. 2016).

[22] Lee, D. (이대우), '북한 무인기: 새로운 비대칭 무기' [North Korean UAVs: A new asymmetric weapon], 정세와 정책 [Situation and policy], May 2014, Sejong Institute.

[23] '北 1m 신형 무인기 첫 확인…"휴전선 도발 우려"' [First confirmation of North Korea's new 1-metre UAV: 'DMZ provocation fear'], KBS News, 18 July 2016; and 'Seoul examines "North Korea drone"', BBC, 2 Apr. 2014.

[24] South Korean Government, '국지방공레이더 연구개발 성공! 북한 소형 무인기까지 탐지 가능해...' [Success in research and development of national radars! North Korean small drones can be detected], Press release, 14 July 2017.

[25] Ahn, J. H., 'What a North Korean drone crash reveals about the country's UAVs', NK News.org, 22 June 2017.

III. The impact on the Korean peninsula

Unmanned systems

One of the key concerns over unmanned systems is the possibility that a nuclear-armed state would use a UAV or UUV to deliver a nuclear weapon as an alternative to an intercontinental ballistic missile (ICBM). Such technology already exists, and North Korea already possesses nuclear warheads.[26] However, technological constraints and international sanctions mean that this development is unlikely to appear in North Korea in the near future. While North Korea has found the means to circumvent financial constraints through cyber means, sanctions have made it difficult for it to import the necessary components and technologies.

In contrast, South Korea has the economic and technical foundations that better enable integration of AI technology into its unmanned systems. One development in South Korea that may affect nuclear risks on the Korean peninsula is Dronebot Jeontudan. The primary purpose of this unit is to carry out reconnaissance tasks targeting North Korea's nuclear and missile sites. It could also launch swarm attacks in the event of a conflict.[27] Depending on the actual effect when the unit operates, this has the potential to undermine the nuclear deterrence capabilities of North Korea.

Cyber operations

Cyber operations are unique in that they allow North Korea the geographic scope and reach to target the USA, which has historically provided South Korea with protection via extended deterrence under its nuclear umbrella. Compared to South Korea's more restrained cyber operations, which are mostly defensive in nature, North Korea appears more willing to use cyber instruments for offensive operations.

AI-enabled cyberattacks can facilitate identification by North Korea of zero-day vulnerabilities in South Korean and US computer systems. In this case, nuclear command, control and communications (NC3) systems may be compromised and US extended nuclear deterrence on behalf of South Korea may lose its effectiveness.[28] Meanwhile, there has been discussion in the USA on the option of 'left of launch' cyberattacks that would defeat the nuclear threats posed by North Korea before it was able to launch a nuclear-armed missile.[29] Such activities suggest that more thought needs to be given to the AI-enabled cyber dimensions of deterrence.

[26] Mizokami, K., 'What's in North Korea's drone arsenal', *Popular Mechanics*, 22 Jan. 2016.; and Nuclear Threat Initiative, 'North Korea', accessed 21 Aug. 2019.

[27] 'South Korea to create "drone-bot combat unit" to swarm North', *Financial Times*, 6 Dec. 2017; and 'S. Korean Army to form weaponized drone unit next year', Yonhap News, 5 Dec. 2017.

[28] Avin, S. and Amadae, S. M., 'Autonomy and machine learning at the interface of nuclear weapons, computers and people', ed. V. Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. I, *Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019), pp. 105–18, p. 107.

[29] Ellison, R., 'Left of launch', Missile Defense Advocacy Alliance, 16 Mar. 2015; and Panda, A., 'The right way to manage a nuclear North Korea', *Foreign Affairs*, 19 Nov. 2018.

IV. Conclusions

Many of the developments in applications of AI to unmanned systems and cyber operations in both North Korea and South Korea remain underway. It is difficult to accurately assess their possible impacts at this stage.

It is evident that both countries are keen to integrate AI into their military weapon systems. However, the implications for nuclear risk remain limited due to technological constraints. This, however, does not rule out possible advances and even leapfrogging of these two countries in the military use of AI. In particular, some dual-use AI technology may proliferate to North Korea to boost its military capabilities and this remains a trend to watch.

7. Artificial intelligence and military advances in Russia

VASILY KASHIN*

Artificial intelligence (AI) technology, in particular machine learning, is among the key priorities of the Russian Government.[1] As such, investment in research and development in both the military and civilian sectors is likely to enjoy a privileged position for the foreseeable future. This essay gives an overview of Russia's advances in AI, in particular the military applications. It begins (in section I) with a review of the defence policies and economic rationale of the Russian Government, explaining the centrality of AI within its overall development model. It then (in section II) explores how the Russian military is applying AI in its military systems, highlighting advances in robotics and unmanned systems that have implications for both combat and nuclear delivery. The essay concludes (in section III) with an assessment of the future of these systems, anticipating greater cooperation with China in the future.

I. Russian defence policies and economic foundations for AI

Russia's defence policies are straightforward, being designed to maintain the credibility of its deterrence capabilities and the competitiveness of its domestic defence industries. The challenge lies in the fact that Russia must do this in the face of its own limited financial means, in contrast to the political adversaries and industrial competitors that possess far greater resources. The main competitor for Russia's defence technology industry is China, which basically targets the same export markets as Russian arms producers (India and Viet Nam being two notable exceptions). Russia's main political adversary is the United States since the two countries have openly identified each other as a major source of military threat. Political relations between the two are expected to remain openly hostile for decades.

Russia has a much smaller economy than its peers: its purchasing power parity-based gross domestic product (GDP) in 2018 was just 15.7 per cent of China's and 19.4 per cent of the USA's.[2] Russia has also been slowly decreasing its military spending as a share of GDP, and it is not expected to rise in the near future.[3] The only way to remain competitive in such an environment is to focus available

[1] Daws, R., 'Putin outlines Russia's national AI strategy priorities', AI News, 31 May 2019.

[2] World Bank Open Data, <https://data.worldbank.org/indicator/NY.GDP.MKTP.PP.CD?locations=RU-CN-US>.

[3] Radin, A. et. al., *The Future of the Russian Military: Russia's Ground Combat Capabilities and Implications for US–Russia Competition* (RAND Corporation: Santa Monica, CA, 2019), p. xii.

* The views expressed in this essay are those of the author and do not necessarily reflect those of any organization to which he is affiliated.

resources on a limited set of priorities in arenas in which decisive technological breakthroughs can be expected, while relatively neglecting the other sectors. Among these key sectors, AI ranks high on the list. Nonetheless, Russia lacks a comprehensive national programme for AI development.

Despite its relatively belated start, in February 2019 Russian President Vladimir Putin ordered that work on such a programme begin. It is currently under development by the Ministry of Digital Development, Communications and Mass Media.[4] Putin highlighted AI in his annual addresses to the Federal Assembly in 2018 and 2019 and has stated that the country that becomes the leader in AI technology will be the 'ruler of the world'.[5]

To this end, AI, big data and machine learning figure prominently in another ongoing major national project: the Digital Economy National Programme, which was approved by the Cabinet of Ministers in July 2017.[6] This programme is aimed at modernization of the general economy, industry and infrastructure, with AI, big data and neural technologies as its core technologies.[7] The approved government funding for this programme for 2019–24 is $32.7 billion.[8] With this foundation, Russia is poised to accelerate its military integration of AI.

II. Military applications of AI

The Russian military is even more enthusiastic about AI than its civilian counterparts. In the military field, AI is one of the major areas of investment under Russia's research agency for advanced military technologies, the Russian Foundation for Advanced Research Projects in the Defence Industry (Fond perspektivnykh issledovanii, FPI)—a functional counterpart to the US Defense Advanced Research Projects Agency (DARPA). The FPI is engaged in the development of AI-related industrial standards for the defence industry and the economy in general.[9]

Russian military analysts expect that the military AI projects that will be implemented most quickly will include (*a*) mathematical modelling of tactical situations to plan operations and to calculate the amount of forces and resources necessary to implement tasks; (*b*) integrated command, control, communications,

[4] President of Russia, 'Перечень поручений по реализации Послания Президента Федеральному Собранию' [List of instructions to implement the presidential address to the Federal Assembly], 27 Feb. 2019; Manz, A., 'Стратегия по искусственному интеллекту появится к концу мая' [An artificial intelligence strategy will appear by the end of May], Politika Segodiya, 7 Mar. 2019; and Bendett, S., 'Russia racing to complete national AI strategy by June 15', Defense One, 14 Mar. 2019.

[5] President of Russia, 'Presidential address to the Federal Assembly', 1 Mar. 2018; President of Russia, 'Presidential address to the Federal Assembly', 20 Feb. 2019; and 'Путин: лидер в сфере искусственного интеллекта станет властелином мира' [Putin: the leader in the field of artificial intelligence will become the ruler of the world], RIA Novosti, 1 Sep. 2017, (author translation).

[6] President of Russia, 'Instructions concerning the implementation of the Digital Economy of the Russian Federation national programme', Russian Federation, 28 Feb. 2019.

[7] 'Программа "Цифровая экономика Российской Федерации"' ['Digital economy of the Russia Federation' programme], Adopted by Russian Government Order no. 1632, 28 July 2017.

[8] Balenko, E. and Posypkina, A., 'Медведев утвердил бюджет национальной программы «Цифровая экономика»' [Medvedev approved the budget of the 'Digital Economy' national programme], RBC, 17 Sep. 2018.

[9] Bendett, S., 'Russia racing to complete national AI strategy by June 15', Defense One, 14 Mar. 2019.

computers, intelligence, surveillance and reconnaissance (C4ISR) systems; (*c*) implementation of AI in remotely operated strike and reconnaissance systems; (*d*) autonomous systems for protection of certain high-value objects; (*e*) battlefield security; and (*f*) simulators and training systems.[10]

Some of the systems from this list are already present in the Russian military. For example, the Nerekhta robotic system—a small autonomous combat vehicle to protect nuclear assets against enemy saboteurs—has already been placed in service by the Russian Strategic Rocket Force.[11] There are also experiments to integrate the system into mechanized infantry units.[12] The new, upgraded version of the Akatsiya-M command-and-control system is reported to have a high degree of autonomy in implementing certain tasks, such as control over execution of commands and analysis of chemical and biological threats.[13]

Russia is one of few countries that has already gained some experience with the use of remotely controlled land robotic systems in actual combat. Uran-9 vehicles—unmanned light tanks with 30-millimetre guns and anti-tank guided missiles—have shown some limitations when operating jointly with other troops. This includes inadequate situational awareness of the robot's operator, which leads to difficulties when interacting with other units.[14] So while the vehicle has been adopted for mass production, it is widely anticipated that its use will be limited. Instead, it is destined to implement such specialized tasks as armed reconnaissance against enemy fortifications and counterterrorist operations. In addition, Russia is seeking to widen the use of AI technology in such vehicles to make them more autonomous. For example, in 2018 the Kalashnikov defence manufacturing company showed a prototype of an autonomous armoured turret capable of independently acquiring, identifying and engaging targets.[15]

Russia also appears to prioritize unmanned underwater vehicle (UUV) technology with a new generation of UUVs tasked with strategic missions. One example is the Poseidon unmanned, nuclear-powered platform with intercontinental range (also known as Status-6), which has been undergoing sea trials.[16] One of its missions will be as a nuclear strategic delivery system, designed to enhance Russia's second-strike capability. Like China, Russia is considering

[10] Burenok, V. M., Durnev, R. A., Krukov, K. U., 'Разумное вооружение: будущее искусственного интеллекта в военном деле' [Reasonable weapons: the future of artificial intelligence in military affairs], *Vooruzheniie i ekonomika*, no. 1(43) (Jan. 2018), pp. 4–13.

[11] 'Боевой робот "Нерехта" охранял "Тополь-М" на учениях под Иркутском' [The combat robot "Nerekhta" guarded "Topol-M" at exercises near Irkutsk], RIA Novosti, 31 Mar. 2016.

[12] Shirokova, I., 'РТК «Нерехта»: результаты испытаний' [RTK 'Nerekhta': test results], *Dyagterevets*, no. 2(10673) (27 Jan. 2017), pp. 2–3.

[13] Solov'eva, N., 'Минобороны закупает мобильные системы управления армией' [Defense Ministry buys mobile army control systems], IT World, 30 June 2018.

[14] Uferev, S., 'Ненадёжный и ненаблюдательный. О недостатках боевого робота «Уран-9»' [Unreliable and unobservant: on the shortcomings of the combat robot 'Uran-9'], *Voennoe Obozrenie*, 21 June 2018.

[15] Peshkov, A., '«Калашников» показал боевой искусственный интеллект в действии' [Kalashnikov showed combat artificial intelligence in action], TV Zvezda, 1 Oct. 2018.

[16] 'Key stage of Poseidon underwater drone trials completed, says Putin', TASS, 2 Feb. 2019; and 'Russia begins testing of "Poseidon" underwater nuclear drone', PressTV, 26 Dec. 2018. On Poseidon see also chapter 12 in this volume.

UUV development as a way to compensate for the overall US advantage in naval warfare. Accordingly, Russia is investing into relevant AI research.

Another area in which the use of the AI is progressing at a fast pace is electronic warfare systems. Among these are the new RB-109A Bylina early-warning system. According to the manufacturer, this platform is capable of operating autonomously, with only the operator providing supervision, to provide air defence.[17] The FPI is also reportedly financing a project to develop a technology to rapidly analyse satellite imagery using machine learning technology.[18] Russian military strategists and practitioners foresee a gradual increase in the role of AI in air combat platforms, which may eventually lead to fully autonomous combat systems that would dominate the sixth generation of combat aircraft. Russia's Okhotnik unmanned combat aerial vehicle (UCAV) is currently undergoing tests and is planned to be the basic platform for applying such an approach.[19]

With these current initiatives, the next stage in the military use of AI will most likely be associated with progress in the creation of human–machine interfaces, as well as a new series of sensors to be included in new AI-controlled systems. New AI-driven design technology, fully autonomous vehicles and a new generation of highly protected communications networks are also on the drawing board. In essence, the development of the AI is expected to engender major advances in every kind of military activity, from combat to medical services.[20]

III. Conclusions

Russia's military and civilian leadership is understandably paying significant attention to the development of AI. It hopes to secure a strong future position for the Russian AI industry by concentrating significant financial and human resources toward this goal. In the current difficult international environment Russia has lost many channels for cooperation with Western partners.

However, potential for cooperation with China is being actively explored and some projects are already being pursued, such as discussions on Huawei acquiring one of Russia's leading facial recognition technology providers.[21] In this manner, Russian and Chinese collaboration in the overall field of AI is only likely to grow.

[17] 'В войска радиоэлектронной борьбы придет искусственный интеллект' [Artificial intelligence will come to electronic warfare troops], Mirovoe obozrenie, 4 Apr. 2017; and Gavrilov, A. and Labunsky, A., 'Искусственный интеллект для ПВО' [Artificial intelligence for air defence], *Arsenal Otechestva*, vol. 35, no. 3 (Mar. 2018).

[18] 'ФПИ создаст технологию дешифровки снимков из космоса с помощью искусственного интеллекта' [FPI will create a technology for decrypting images from space using artificial intelligence], TASS, 18 Jan. 2018.

[19] 'БЛА «Охотник» станет прототипом истребителя следующего поколения' [UAV 'Okhotnik' will become the prototype of the next generation fighter], *Voennoe Obozrenie*, 20 July 2018.

[20] Burenok et al. (note 10).

[21] Polyakova, A., '«Ъ»: Huawei задумалась о покупке российского разработчика систем распознавания лиц «Вокорд»' [Komersant: Huawei is thinking about acquiring the Russian developer of facial recognition systems 'Vocord'], Rusbase, 25 Jan. 2019.

8. Exploring artificial intelligence and unmanned platforms in China

LORA SAALMAN*

China has long maintained ambiguity about the role of unmanned vehicles.[1] This can be contrasted with Russia, which has a clearer aim of employing unmanned vehicles transiting sea, air and space as platforms for nuclear weapons.[2] This essay reviews China's approach to the use of artificial intelligence (AI) and autonomy in unmanned vehicles. It starts (in section I) with an overview of the underlying assumptions that underpin research on AI and autonomy in China. It then (in section II) analyses how these developments are playing out in military advances and rivalries among China, Russia and the United States. The essay concludes by emphasizing the importance of combining analysis of assumptions and technologies when discussing the future of deterrence relations among the three countries (in section III).

I. Assumptions underpinning research on AI and autonomy

A review of 904 Chinese-language technical and strategic articles, papers and books reveals a focus on the integration of AI and autonomy to facilitate fault detection and diagnosis; embedded training systems, simulation and modelling; and data accumulation and processing for remote sensing and situational awareness.[3] None of these activities is necessarily destabilizing in and of itself, even considering the crossover between China's civilian and military research and development. In some respects, improvements in China's reconnaissance capabilities and in the reliability of a range of its platforms could be a stabilizing measure. If China gains greater situational awareness and can ensure its nuclear retaliatory capabilities, then some of its insecurities about an unanticipated first strike may be mitigated. Yet Chinese insecurities are not simply a question of

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

[2] Sayler, K. M., *Hypersonic Weapons: Background and Issues for Congress*, Congressional Research Service (CRS) Report for Congress R45811, (US Congress, CRS: Washington, DC, 11 July 2019); and Peck, M., 'Russia has begun underwater tests of its Poseidon thermonuclear torpedo', *National Interest*, 19 May 2019.

[3] These writings were produced by a wide range of Chinese universities and institutes, including the China Electronics Technology Group, the Academy of Armoured Forces Engineering, the Tactical Weapons Division of the China Academy of Launch Vehicle Technology, the Department of Information Operations and Command Training of the PLA National Defence University, the Naval Aeronautical Engineering Institute of Qingdao, the Department of National Defence Architecture Planning and Environmental Engineering, the Laboratory of Special Fibre Optics and Optical Access Networks, and the National Key Laboratory of Integrated Service Networks and Key Technologies.

* The views expressed in this essay are those of the author and do not necessarily reflect those of any organization to which she is affiliated.

technology. They are also rooted in a set of concerns about false negatives and assumptions about the capabilities and intent of the USA.

To better understand these trends, it is important to review the underlying assumptions of China and the USA as revealed by both official and unofficial documents. In the USA, military analysts are often preoccupied with the concern that alarms or early-warning systems, accidentally or even intentionally triggered, could produce false positives. Chinese analysts, in contrast, are much more concerned with false negatives. In other words, they are preoccupied with the inability of China's systems to identify, much less counter, a stealthy and prompt precision strike. This mirrors misgivings also found in Russia.[4] China's assumptions about its own challenges in early warning, combined with its concerns over US advances in high-precision, high-speed systems—from Conventional Prompt Global Strike (CPGS) to spaceplanes—imply that technologies such as AI and autonomy could take on destabilizing qualities.[5] As China further develops its concept of 'rapid response' (快速反应), as cited in its 2015 military strategy, its integration of AI and autonomy into military systems is likely to increase.[6] This ranges from allegations of automation-enabled launch-on-warning for its missiles to autonomy- and AI-enabled manoeuvrability and precision guidance for hypersonic glide platforms.[7]

While this concept of 'rapid response' is not featured in the 2019 version of China's military strategy, Chinese technical writing reveals that it is still present in practice if not in stated doctrine.[8] These works on AI and autonomy continue to reveal that China has a strong emphasis on unmanned aerial vehicles (UAVs) and unmanned underwater vehicles (UUVs), and in increasing speed and accuracy of response.[9] In fact, Lin Yang, marine technology equipment director at the Shenyang Institute of Automation of the Chinese Academy of Sciences, has reportedly confirmed that China is developing a series of extra-large unmanned underwater vehicles (XLUUVs).[10] This is noteworthy, since the Shenyang Institute of Automation is a major producer of underwater robots for the Chinese military and Lin developed China's first autonomous underwater vehicle with operational

[4] Podvig, P., 'Russia and the Prompt Global Strike plan', PONARS Policy Memo no. 417, PONARS Eurasia, Dec. 2006.

[5] Luo, G. (罗桂兰), Cheng, M. (成茂荣) and Shi, W. (石文斌), '以国家安全战略需求为牵引加快推进战略预警系统建设' [Accelerating construction of a strategic early warning system based on national security strategy demands], 国防科技 [National Defense Science & Technology], vol. 33, no. 6 (June 2012). On CPGS see chapters 13 and 14 in this volume.

[6] Chinese State Council, 中国的军事战略 [China's military strategy], White paper (State Council Information Office: Beijing, May 2015).

[7] Acton, J. (ed.), *Entanglement: Russian and Chinese Perspectives on Non-nuclear Weapons and Nuclear Risks* (Carnegie Endowment for International Peace: Washington, DC, 2017); and Saalman, L., 'Prompt global strike: China and the spear', Independent faculty article, Asia–Pacific Center for Security Studies, Apr. 2014.

[8] Chinese State Council, 新时代的中国国防 [China's national defence in the new era], White paper (State Council Information Office: Beijing, July 2019).

[9] This assessment is based on the author's forthcoming work and research into 400 new Chinese-language technical writings.

[10] Glass, P., 'China's robot subs will lean heavily on AI: report', Defense One, 23 July 2018; and Chen, S., 'China military develops robotic submarines to launch a new era of sea power', *South China Morning Post*, 22 July 2018.

depth beyond 6 kilometres. He is now chief scientist of the 912 Project, a classified programme that is reportedly developing new-generation military underwater robots in time for the 100th anniversary of the Chinese Communist Party, in 2021.[11]

These types of unmanned platforms, ranging from small to large, have a range of potential applications cited throughout Chinese works, including enhanced accuracy in battlefield reconnaissance, surveillance, patrolling, electronic reconnaissance and communications; electronic interference, combat assessment and radar deception; projectile firearms, laser guidance, target indication and precision bombing; intercept and launch of tactical missiles and cruise missiles; anti-armour, anti-radiation and anti-naval capabilities; and nuclear, chemical and biological detection and operations. When leveraging new means of warfare, Chinese experts also discuss the use of swarm systems for a number of purposes, with battlefield applications focusing on anti-submarine warfare and countering integrated air defence.[12]

AI and autonomy are means of providing China with an opportunity to exploit a new technological niche of excellence, but they are not ends in and of themselves. This is one of many reasons why China's leadership has had misgivings regarding arms control for autonomous systems. Moreover, the amount of Chinese research already being conducted in this field, particularly at the university level, is substantial and unlikely to diminish in the short term. Programmes on AI and autonomy receive ample government support through such funds as the Laboratory of National Defence Technology for Underwater Vehicles, the Project for National Key Laboratory of Underwater Information Processing and Control, the National Key Basic Research and Development Programme, the China Aviation Science Foundation, the National Science and Technology Major Project, the National 973 Project, the National Key Laboratory Fund, the National 863 High-tech Research and Development Programme, and the Ministry of Communications Applied Basic Research Project, among a number of others.[13]

II. Military applications of AI and autonomy

Expansive Chinese programmes to integrate AI and autonomy in weapon systems, even in such challenging or hypothetical domains as underwater swarms, indicate the emphasis that this research receives within the hierarchy of national defence planning. Whether or not China is able to achieve all of these capabilities, China's New Generation Artificial Intelligence Development Plan indicates the vast scale

[11] Glass (note 10).

[12] An, M. (安梅岩), Wang, Z. (王兆魁) and Zhang, Y. (张育林), '人工智能集群控制演示验证系统' [Demonstration and verification system for artificial intelligent swarm control', 机器人 [Robot], vol. 38, no. 3 (Mar. 2016), pp. 265–75. See also chapter 4 in this volume.

[13] 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*, Strategic Multilayer Assessment (SMA) Periodic Publication (Department of Defense: Washington, DC, Dec. 2018), pp. 153–60; and Saalman, L., 'Factoring Russia into the US–Chinese equation on hypersonic glide vehicles', SIPRI Insights on Peace and Security no. 2017/1, Jan. 2017.

of domestic priorities and investments in AI. Given this document's foundation in 'military-civilian fusion' (军民融合), it merits a thorough analysis of AI dual-use applications and comparison with Chinese technical journals to determine what has been achieved and what capabilities are likely to appear soon.[14] The direct implications of aerial and underwater swarms for larger, more lumbering US nuclear and conventional platforms remain to be seen. However, if the USA proceeds with low-yield submarine-launched ballistic and cruise missiles, as proposed in its 2018 Nuclear Posture Review, then China could deploy swarms to track and potentially intercept US dual-capable platforms.[15] In short, whether intentional or unintentional, an escalatory scenario could develop.

The concept of using nuclear coercion or force to pre-emptively de-escalate a conventional conflict—as is implied by the 2018 US Nuclear Posture Review—cuts to the core of China's concerns over US nuclear coercion and intentional escalation. Even if interpreted as a US response to perceived Russian postural shifts, there will also be an impact on China.[16] The issue is not limited to lowering the threshold for nuclear use—this US shift further eradicates the taboo against nuclear use.[17] For China, which has been expanding its nuclear arsenal at a relatively modest pace, the prospect of the USA resuming a forward-deployed, tactical nuclear posture exacerbates its sense of encirclement. Such a posture also amplifies China's perceived and real vulnerability to US ambitions to field both kinetic and surveillance platforms such as the X-37B orbital test vehicle and the X-43A and X-51 hypersonic vehicles, among others.

The evolution of smaller unmanned platforms mobilized in joint formations could turn China's nuclear asymmetrical disadvantage on its head. Much like decoys, which can be used as an inexpensive means of confusing and saturating missile defences, low-cost swarms of unmanned aerial, water and ground vehicles along with cyber technologies could provide a guerrilla combat-style advantage against systems that the USA sees as providing an element of surprise, speed and precision. Some of these platforms are already destined for deployment and will provide China with greater capability to monitor US activities in the Asia–Pacific region. However, if these platforms are adapted for combat—in efforts to disrupt or confront lower-yield, smaller-scale US nuclear or dual-capable platforms—the potential for miscalculation may grow.

If China enhances its development of cruise missiles and hypersonic glide platforms by applying AI and autonomy, close-range encounters off the coast of Taiwan and in the East China and South China seas could grow even more complicated. China's ground-launched DH-10 missile is believed to carry a conventional

[14] Chinese State Council, '新一代人工智能发展规划' [New Generation Artificial Intelligence Development Plan], Order no. 35, 8 July 2017; Miracola, S., 'Beijing's ultimate goal: the military-civilian fusion', Italian Institute for International Political Studies, 3 Aug. 2018; and Saalman (note 1).

[15] US Department of Defense (DOD), *Nuclear Posture Review* (DOD: Washington, DC, Feb. 2018), pp. 54–55.

[16] Schneider, M. B., 'Escalate to de-escalate', *Proceedings* (US Naval Institute), vol. 143, no. 2 (Feb. 2017).

[17] Li, B. and Zhao, T. (eds), *Understanding Chinese Nuclear Thinking* (Carnegie Endowment for International Peace: Washington, DC, 2016); and Li, B., 'Chinese thinking on nuclear weapons', *Arms Control Today*, vol. 45, no. 10 (Dec. 2015).

warhead, but there are indications that the air-launched CJ-10 missile may have both nuclear and conventional variants.[18] Among other platforms cited as potentially dual capable are China's intermediate-range mobile ballistic missiles DF-26 and H-6K.[19] Moreover, China has hedged on what kind of payload will be carried by its hypersonic glide platforms, such as the DF-ZF, which is designed to break through US defences.[20]

III. Conclusions

With the release of the US Nuclear Posture Review and Russian President Vladimir Putin's subsequent declaration that Russia had developed new nuclear weapons, Russia and the USA have engaged in a game of tit-for-tat.[21] If China also acts in this way, a new set of destabilizing variables could be introduced into a region that is already tense and crowded, with freedom-of-navigation operations carried out among competing territorial claims. The risk of aerial or maritime collision is already high and is likely to be exacerbated.

China's work on integration of swarms that could be used to confront conventional- and nuclear-capable US platforms could result in greater likelihood of collision and confrontation in the air and at sea. Moreover, alleged Chinese discussions on integration of launch-on-warning for missiles and hedging on conventional versus nuclear payloads on hypersonic vehicles create ambiguity—this may be stabilizing in staying the hand of the USA, but could also result in greater escalation, as seen with the most recent US Nuclear Posture Review. With greater application of AI and autonomy in prompt and high-precision systems such as cruise missiles, spaceplanes and hypersonic glide platforms, such ambiguities and their underlying assumptions merit greater attention.

These activities suggest that China's concerns about false negatives could lead to greater automation and autonomy injected into command-and-control operations that run the risk of producing a false positive. Furthermore, Chinese discussions about keeping a human in the loop in technical writings remain limited to non-existent.[22] This indicates a gap in the current discourse, neglecting the potential adverse impact of AI and autonomy on military command and control. Therefore, it is all the more important to understand China's strategic assumptions about false negatives, intentional escalation and rapid response. Exploring these concepts and their technical applications is crucial for gauging how China may integrate AI and autonomy into its conventional and nuclear platforms and how other countries may respond.

[18] 'DH-10 / CH-10 / CJ-10 Land-Attack Cruise Missiles (LACM) Hong Niao / Chang Feng / Dong Hai-10', GlobalSecurity.org, accessed 22 Apr. 2019.

[19] Office of the Secretary of Defense, *Annual Report to Congress: Military and Security Developments Involving the People's Republic of China 2019* (Department of Defense: Washington, DC, 3 May 2019), p. 41.

[20] Saalman, 'Factoring Russia into the US–Chinese equation on hypersonic glide vehicles' (note 13).

[21] 'Russia's Putin unveils "invincible" nuclear weapons', BBC, 1 Mar. 2018.

[22] Ni, J. (倪建军) et al., '关于无人机数据链系统智能化问题的思考' [On the intellectual problem of UAV data links], 第五届中国指挥控制大会论文集 [Proceedings of the 5th China Command and Control Conference] (China Command and Control Society: Beijing, 2017).

Part II. The future of arms control and strategic stability with artificial intelligence

This part builds on the technologies and dynamics featured in part I to explore how artificial intelligence (AI) is transforming strategic stability and arms control in East Asia. Strategic stability is no longer simply the purview of nuclear-armed states and their allies. Instead, the authors of the following essays maintain that strategic stability will be increasingly driven by AI advances that promise to reshape technological asymmetries and to challenge the underpinnings of existing arms control structures.

These essays update and expand on historical definitions and regulatory frameworks to formulate a more responsive and adaptive set of confidence-building measures (CBMs). Recognizing the difficulty of controlling a sphere as inherently dual-use as AI, this part provides the reader with a foundation for addressing the future of AI integration into nuclear forces within such existing structures as the 1968 Non-Proliferation Treaty and the 1980 Convention on Certain Conventional Weapons, among others.

In the first two essays, Jiang Tianjiao (chapter 9) and Cai Cuihong (chapter 10) systematically lay out a framework for understanding how AI is contributing to an increasingly complex set of strategic stability and deterrence relations. In the first essay, Jiang contextualizes AI-related threats to stability in terms of nuclear deterrence, proliferation and terrorism; laws and norms; and human–machine interaction and conflict. Cai furthers this overview by mapping how various AI-related applications are shaping strategic stability through power dynamics, threat perceptions and nuclear postures. In combination with Jiang's overview, Cai's essay provides a conceptual layout for part II by analysing AI's enhancement of nuclear and conventional weapons, as well as the behavioural risks and psychological anxieties that this creates among countries.

The second pair of essays details various threats posed by AI and unmanned systems to both stability dynamics and the physical environment, along with means of addressing these challenges via controls. Vadim Kozyulin (in chapter 11) explores how AI in lethal autonomous weapon systems (LAWS) and command, control, communications, computers, intelligence, surveillance and reconnaissance (C4ISR) systems can exacerbate strategic time pressure. In confronting these risks, he suggests means of enhancing existing regulatory bodies and mechanisms. Carrying these concerns further, Hwang Il-Soon and Kim Ji-Sun (in chapter 12) use the anticipated environmental damage from nuclear warhead and nuclear reactor fallout from the Poseidon unmanned underwater vehicle (UUV) as a litmus test for the effectiveness and responsiveness of the nuclear non-proliferation regime in addressing emerging threats.

The third set of essays details AI applications in nuclear forces and arms control. Arie Koichi (in chapter 13) first discusses how Chinese, Russian and US military advances are having an impact on East Asian regional dynamics. He notes that

current discussions of the Poseidon UUV and AI-enhanced nuclear submarines remain at the operational or tactical level, with limited or no understanding at the strategic level. He recommends that, to avoid a cascading series of unexplored nuclear risks, more scenario-building exercises are essential. Nishida Michiru (in chapter 14) then expands future CBM options by discussing how AI fits into and departs from traditional arms control. Highlighting the dual-use nature of AI technology, he concentrates on behaviour-based approaches and CBMs as a way forward.

LORA SAALMAN

9. The impact of military artificial intelligence on warfare

JIANG TIANJIAO*

Chinese scholars have paid close attention to the application of artificial intelligence (AI) in the military field and its nature of serving as a double-edged sword. On the positive side, machine learning technology can improve the performance of weapon systems and strengthen command-and-control systems. Intelligent weapons and decision-support systems will greatly enhance the efficiency of future wars, enabling the potential for hyper-speed warfare. However, military AI applications will also have adverse effects on overall world peace and stability.

This essay details these impacts and how they may be addressed. It starts (in section I) with a discussion of the human costs of war, which are likely to decrease with the introduction of unmanned systems while, conversely, increasing the willingness of countries to engage in conflict. It then considers (in section II) how laws and norms are likely to be re-shaped and raises questions about the dynamics between humans and machines in warfare. It concludes (in section III) by offering recommendations on how to address the offence–defence imbalance and instability that may result from AI advances.

I. Costs and thresholds

The combination of AI and unmanned combat platforms has led to a significant decline in the cost of future wars.[1] As a result, there is destined to be a gradual decline in the threshold of war and an increase in belligerent tendencies among states. One of the main reasons why war is less frequent and less intense today is that, after the two world wars, international regulations under the United Nations Charter advocated for peaceful settlement of international disputes and prohibited acts of war, unless in self-defence.[2]

This longing for peace stemmed from the fact that mankind experienced painful sacrifices and suffering during the two world wars. However, with the rise of AI and unmanned combat platforms, future wars may no longer cost soldiers' blood, much less their lives. Once unmanned combat platforms are mass-produced, the cost of war will decrease. Launching war will not only no longer confront political

[1] Scharre, P. and Burg, D., 'To save money, go unmanned', War on the Rocks, 22 Oct. 2014; Lewis, M. W., 'Drones: actually the most humane form of warfare ever', *The Atlantic*, 21 Aug. 2013; and Walsh, J. I. and Schulzke, M., *The Ethics of Drone Strikes: Does Reducing the Cost of Conflict Encourage War?* (US Army War College, Strategic Studies Institute: Carlisle, PA, Sep. 2015).

[2] Charter of the United Nations, signed 26 June 1945, entered into force 24 Oct. 1945.

* The views expressed in this essay are those of the author and do not necessarily reflect those of any organization to which he is affiliated. It was translated from Chinese to English by the volume editor, Lora Saalman.

and legal taboos but will also be likely to stimulate relevant industrial chains to make AI-enabled combat more profitable.

A core important reason why war has not erupted among major powers from the cold war until today has been the existence of nuclear weapons. However, AI and its military applications are likely to engender conflict as it undermines nuclear strategic stability. This is particularly the case when AI technology is combined with remote sensing and unmanned aerial vehicle (UAV) surveillance, which assist in more accurate detection of an opponent's deployment of strategic power. This reduces strategic stability, which is predicated on a certain level of uncertainty. Additionally, some autonomous weapon platforms, including unmanned underwater vehicles (UUVs), increase the risks of accidental launch and nuclear war. They do this through the potential for accidents and lack of human control.

The proliferation of AI and related unmanned combat platforms also increases the risk of terrorism. It will be difficult for the international community to ensure that relevant technologies do not fall into the hands of terrorists. Related equipment, including UAVs, is also likely to spread if it accidentally crashes or is shot down and captured. Especially when these unmanned devices are equipped with weapon systems, including low-yield nuclear weapons, this proliferation can elicit global disasters.

II. Laws, norms and ethics

Similar to cyberwarfare, AI represents a new technological revolution and has no corresponding concepts and definitions in international law and norms. For example, there remain serious challenges as to whether and how the traditional key concepts of the law of war—such as distinction between civilians and combatants and principles of proportionality in armed conflict—can be applied to AI and related unmanned combat platforms.

The second edition of the Tallinn Manual on cyberwarfare, compiled by legal scholars in the international community, offers some guidance.[3] However, the discussion of AI laws and regulations needs to be further strengthened. Before the norm on the military use of AI is defined, it cannot be ruled out that countries will use loopholes and grey areas in the rules to undermine regional peace and stability.

The military application of AI will eventually lead to ethical issues of human–machine conflict. Should humanity place its destiny in the hands of machines that are more intelligent than human? Can machines replace humans in decision-making and even determine the ultimate direction of human civilization? Once AI is widely used in the military field, these dilemmas will become unavoidable.

[3] Schmitt, M. N. (ed.), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press: Cambridge, 2017).

III. Conclusions

Despite these adverse effects, the current development of AI technology is still in its infancy. There are many cases of use that can be envisaged, but there are few that are actually available for research and analysis. In fact, much as with the cyber technology revolution, the subversive role of AI at the military and strategic levels is likely to have been exaggerated. China's traditional strategic culture emphasizes that pure technological revolution alone cannot determine the outcome of war: the core factor remains the role of human beings and the political intentions behind them.

For example, with the combination of AI and deterrence strategy, what is the ultimate goal? If it is simply to defeat an opponent, then AI technology will undoubtedly increase the efficiency of warfare. However, if the strategic goal is to deter opponents from acting rashly or to force the opponent to act in a certain way, then AI may not be able to play a huge role. In the 1980s, scholars of international relations, including Robert Jervis and Richard Ned Lebow, criticized deterrence theory, arguing that so-called 'rationality' only existed in an ideal state.[4] In reality, leaders face too many complexities in decision-making, including historical, cultural, ethnic, religious and other factors, leading to cognitive bias. Combined with the competing interests of bureaucratic departments, these factors often lead to an apparently irrational decision. In this complex decision-making system, it is questionable how much of a role AI can play.

Nonetheless, AI naturally conforms to the overall laws of interaction between such a technological revolution and international relations. Such technological revolutions have occurred in the past with nuclear and cyber advances. The emergence of a new technological revolution often starts with competition among countries, especially the major powers. According to this logic, as soon as the new technology is mastered, the offence–defence balance will be broken and a state will be able to obtain greater power. However, historical experience shows that the spread of technology is often much faster than that anticipated by major powers. When unable to monopolize revolutionary technology, countries will engage in an arms race until non-proliferation becomes in their common interest. Thus, it is imperative to engage in early reflection on and criticism of new technologies in order to assist the international community to form a new set of norms and legal frameworks. Once AI technology reaches a mature level of development and technological diffusion, this foundation can then be used to usher in greater reflection and restraint.

[4] Jervis, R., 'Rational deterrence: theory and evidence', *World Politics*, vol. 41, no. 2 (Jan. 1989), pp. 183–207; and Lebow, R. N. and Stein, J. G., 'Rational deterrence theory: I think, therefore I deter', *World Politics*, vol. 41, no. 2 (Jan. 1989), pp. 208–24.

10. The shaping of strategic stability by artificial intelligence

CAI CUIHONG*

The world has already begun to enter the artificial intelligence (AI) era. AI and unmanned vehicles have been called the 'second nuclear weapon' with the potential to change the ways in which future wars will be fought.[1] China, Russia and the United States, among other powers, have been competing in AI development. The world is thus embarking upon, or perhaps could be said to have already started, a new cold war, this time driven by AI.

In the light of these developments, this essay considers whether AI will have a similarly profound impact on the strategic stability of the great powers. It begins (in section I) with a review of national AI strategies. It then describes (in section II) how the nuclear strategic stability of the cold war has developed into modern complex strategic stability. The essay then considers the conditions under which AI could have an impact on strategic stability (in section III) and what forms this impact could take (in section IV). It ends by considering (in section V) how AI needs to be included in any framework for maintaining strategic stability.

I. National AI strategies

In recent years the US Government has issued a series of documents on AI strategy.[2] Throughout these documents, the USA emphasizes the use of technological innovation to preserve US military advantage into the future—known as the Third Offset Strategy.[3] Moreover, these documents note that no other technology would have as much of an impact on US military operations as AI and intelligent technologies, whether used in remote sensing, command-and-control networks,

[1] '日媒称日本正加快引入"第二核武器" 紧追中美俄步伐' [Japanese media says that Japan is accelerating the introduction of the 'second nuclear weapon' and closely following the pace of China, the United States and Russia], 参考消息 [Reference News], 28 Jan. 2019.

[2] US National Science and Technology Council (NSTC), Networking and Information Technology Research and Development Subcommittee, *The National Artificial Intelligence Research and Development Strategic Plan* (White House: Washington, DC, Oct. 2016); US Executive Office of the President and National Science and Technology Council (NSTC) Committee on Technology, *Preparing for the Future of the Artificial Intelligence* (White House: Washington, DC, Oct. 2016); US Executive Office of the President, *Artificial Intelligence, Automation and the Economy* (White House: Washington, DC, Dec. 2016); US Department of Defense (DOD), *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity* (DOD: Washington, DC, Feb. 2019); and 'Maintaining American leadership in artificial intelligence', Executive Order no. 13 859, 11 Feb. 2019.

[3] Hagel, C., US Secretary of Defense, Keynote speech, Reagan National Defense Forum, 15 Nov. 2014.

* The views expressed in this essay are those of the author and do not necessarily reflect those of any organization to which she is affiliated. It was translated from Chinese to English by the volume editor, Lora Saalman.

operations or logistical support networks.[4] Reflecting these priorities, in June 2017 the US Government considered limiting China's investment in AI in the USA.[5] In doing so, in accordance with the characteristics and advantages of AI technology, the US military sought to take the lead in proposing a new operational concept of algorithmic warfare with machine learning and deep learning technology as its core.

While China's AI developments started late, they also developed rapidly. China has already become an important force leading global innovation and development of AI. In May 2016 a number of Chinese ministries and agencies established the 'Internet Plus' three-year AI action plan to guide AI technological innovation and industrial development.[6] In July 2017 the State Council issued the New Generation Artificial Intelligence Development Plan, which details medium- and long-term systematic deployment of China's AI development.[7] Use of AI has become a national development strategy and the Chinese Government has been increasing financial and policy support.

According to Russian President Vladimir Putin, AI is 'the future, not only for Russia, but for all humankind' and 'whoever becomes leader in this sphere will become the ruler of the world'.[8] So while the scale of Russia's AI industry and its overall development of AI have lagged behind China and the USA, its achievements in certain areas should not be discounted. The Russian military is currently applying AI to its equipment-renewal programme: a three-dimensional intelligent equipment system has gradually formed, encompassing unmanned ground vehicles, unmanned aerial vehicles (UAVs) and unmanned underwater vehicles (UUVs).[9]

Other technological powers have also joined the AI development race with their own scientific and technological strengths. The Japanese Government has proposed a plan for a super-smart society, the Society 5.0 strategy.[10] The British Government released a report on Artificial Intelligence: Opportunities and Implications for the Future of Decision Making in 2016 and an AI 'sector deal' between

[4] Liao, K. (廖凯), '透视美军抵消战略的变与不变' [The changing and unchanging perspective of the US Third Offset Strategy], 解放军报 [PLA Daily], 5 Sep. 2017, p. 7.

[5] Stewart, P., 'US weighs restricting Chinese investment in artificial intelligence', 14 June 2017, Reuters.

[6] Chinese National Development and Reform Commission, Ministry of Science and Technology, Ministry of Industry and Information Technology, and Central Cyberspace Affairs Office Commission, '"互联网+"人工智能三年行动实施方案' ['Internet Plus' artificial intelligence three-year action plan], 18 May 2016.

[7] Chinese State Council, '新一代人工智能发展规划' [New Generation Artificial Intelligence Development Plan], Order no. 35, 8 July 2017.

[8] '普京大帝谈AI：得人工智能者得天下' [Putin the Great discusses AI: getting AI means getting the world], 搜狐网 [Sohu.com], 4 Sep. 2017, (author translation); and '普京警告：发展AI最成功国家将统治全世界，未来是无人机的战争' [Putin warns: countries that are most successful at developing AI will rule the world, drone wars are the future], 搜狐网 [Sohu.com], 3 Sep. 2017; and '"Whoever leads in AI will rule the world": Putin to Russian children on Knowledge Day', RT, 1 Sep. 2017.

[9] Wang, H. (王慧婕), '发展人工智能已成全球之势' [Developing AI has become a global trend], 人民论坛 [People's Tribune], Jan. 2018, pp. 20–21. See also chapter 7 in this volume.

[10] Japanese Cabinet Office, 'Society 5.0', accessed 26 Apr. 2019.

the government and the British AI sector in 2018.[11] France has also striven to become a European leader in AI, with the government launching the country's national AI strategy in 2017 and publishing a vision of 'AI for Humanity' in 2018.[12] Germany's 'Industry 4.0' strategy includes machine perception, planning, policy and human–machine interaction among the key research directions of its AI development.[13]

II. From nuclear strategic stability to complex strategic stability

Strategic stability is a concept from the cold war era. Its general definition is primarily derived from a 1990 Soviet–US joint statement on non-proliferation and strategic stability.[14] According to this statement, 'strategic stability' may be understood as an equilibrium of strategic forces between the Soviet Union and the USA. In other words, the strategic relationship between the two major powers is such that neither side has the motivation to launch a first nuclear strike.[15] The concept of strategic stability born in the cold war period has two components: crisis stability and arms race stability. Its direct purpose was to use the structure of armaments to eliminate the possibility of a nuclear war between the two superpowers. This theory came to be the main foundation of Soviet and US nuclear strategy, guiding mutually assured destruction (MAD) and having an impact on the development of the two countries' strategic nuclear forces throughout the cold war. Although the concept of strategic stability encountered certain challenges in the post-cold war era, it remains the basis for influencing the balance of international strategic forces.

Since the end of the cold war, the Soviet–US bipolar structure that guided the international security environment has undergone tremendous changes. Many Chinese and foreign scholars quickly concluded that the concept of strategic stability was no longer applicable to the new international situation. However, the concept continues to develop. Strategic stability had been limited to a relationship in which there is a lack of opportunity or motivation to destroy all the nuclear forces of the opponent.[16] Russian experts tend to divide this into a narrow and

[11] Innovate UK, 'Artificial Intelligence 2020 National Strategy', Gov.uk blog, accessed 26 Apr. 2019; British Government, *Industrial Strategy: Artificial Intelligence Sector Deal* (Department for Business, Energy and Industrial Strategy: London, 2018); and Jin, D. (ed.), *Reconstructing Our Orders: Artificial Intelligence and Human Society* (Shanghai University Press/Springer: Shanghai/New York, 2018).

[12] French Government, '#FranceIA: the national artificial intelligence strategy is underway', 26 Jan. 2017; and Villani, C., *For a Meaningful Artificial Intelligence: Toward a French and European Strategy* (Conseil national du numérique: Paris, Mar. 2018).

[13] Wang (note 9).

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

[15] Wu, T. (吴艇), '从中美战略稳定性看太空武器化问题' [Examining space weaponization via Chinese–US strategic stability], Master's thesis, Fudan University, Apr. 2012, p. 16.

[16] Logan, J., *China's Space Programme: Options for US–China Cooperation*, Congressional Research Service (CRS) Report for Congress RS22777 (US Congress, CRS: Washington, DC, 29 Sep. 2008); and Colby, E. A. and Gerson, M. S. (eds), *Strategic Stability: Contending Interpretations* (US Army War College, Strategic Studies Institute: Carlisle, PA, 2013).

broad sense.[17] In the narrow sense, strategic stability refers to a state in which military strengths and the potentials of strategic forces are roughly equal and neither side seeks to change the military balance to acquire a sustained advantage. In the broad sense, strategic stability refers to the cumulative implementation by two countries or alliances of political, economic, military and other measures that make it impossible for either party to launch a military offensive. In other words, strategic stability may be narrowly characterized as the balance between major powers, in particular the balance of strength and capabilities of strategic weapons. More broadly, it may be defined as a condition in which global actors maintain self- and mutual restraint on a global scale, thereby engendering a relatively stable and balanced strategic situation within the international system.[18]

As noted in a joint statement issued by China and Russia in June 2016, the international community is accustomed to regarding 'strategic stability' as a purely military concept in the field of nuclear weapons. This does not reflect the broad and multifaceted nature of contemporary strategic issues. To achieve the goal of peace and security, strategic stability should be evaluated from a more comprehensive perspective.[19]

Of course, this kind of strategic stability does not mean that disagreements do not occur. However, these differences should not affect the development of overall relations. As such, it could be argued that nuclear strategic stability during the cold war period has developed into the complex strategic stability of today, which is a comprehensive strategic balance in which both the scope and the subject are diversified and intertwined. In transitioning from the narrow to the broad concept of strategic stability, there have been two important changes, as detailed below.

First, the scope of strategic stability has expanded from nuclear power relations via military and security relations to overall strategic relations. The core of maintaining strategic stability is the achievement of mutual deterrence. For this reason, the concept of cross-domain deterrence has begun to replace the concept of nuclear deterrence among decision makers. In recent years, the USA has been committed to creating a system of strategic deterrence that gives it a dominant global role. At the same time, it is also gradually adjusting this system of strategic deterrence at the cognitive and operational levels. At the cognitive level, the USA's greatest threat has transformed from nuclear terrorism to strategic competition and cross-domain threats. At the operational level, the means of cross-domain deterrence have been strengthened across various fields: to reshape the USA's absolute superiority in nuclear deterrence, to establish offensive and defensive conventional deterrence, and to improve its offensive emerging capabilities in

[17] Dvorkin, V., 'Preserving strategic stability amid US–Russian confrontation', Carnegie Moscow Center, Feb. 2019; Berls, R. E. and Ratz, L., *Rising Nuclear Dangers: Assessing the Risk of Nuclear Use in the Euro-Atlantic Region* (Nuclear Threat Initiative: Washington, DC, Oct. 2015); and Margoev, A., *Pursuing Enhanced Strategic Stability through Russia–US Dialogue* (PIR Center: Moscow, May 2019).

[18] Li, Z. (李喆), '"第二核时代"战略稳定性研究' [Study on strategic stability in the 'second nuclear age'], 江南社会学院学报 [Journal of Jiangnan Social University], vol. 17, no. 4 (Apr. 2015), pp. 32–36, p. 32.

[19] 中华人民共和国主席和俄罗斯联邦总统关于加强全球战略稳定的联合声明 [Joint Statement by the President of the People's Republic of China and the President of the Russian Federation on Strengthening Global Strategic Stability], Beijing, 25 June 2016, (author translation).

cyberspace and space.[20] This is being done with the aim of achieving complementary and flexible combinations of advantages among these various deterrents. Furthermore, advanced AI systems can provide deterrence against potential threats, just like the nuclear weapons of the cold war.

Second, the protagonists of strategic stability have expanded from the two major coalitions led by the USA and the USSR to include various global actors. During the cold war, the paramount figures in strategic stability were the two nuclear superpowers, the USA and the USSR, which gave strategic stability certain characteristics. Since the global power game at that time was highly concentrated on the two superpowers, it was difficult for any third-party forces to influence the power balance between the two camps. As a result, strategic stability equated with the dynamics between the two. During the long period that followed the end of the cold war, the focus of global strategic stability also remained the bilateral strategic stability between the two nuclear superpowers, Russia and the USA.

As the world enters the next nuclear era, however, the issue of strategic stability is no longer limited to strategic nuclear confrontation between two militaries. In the global nuclear power system, it is no longer just two nuclear superpowers that can influence and play a decisive role. Furthermore, countries with strategic nuclear power are no longer limited to the five defined as nuclear weapon states by the 1968 Non-Proliferation Treaty (NPT).[21] In fact, many conventional weapons can already replace some of the functions of nuclear weapons.[22] With the deepening of globalization, the nuclear environment is becoming more and more fractured. Within this complex environment, more actors can influence global strategic stability through such high-technology asymmetric means as AI.

III. The feasibility of AI having an impact on strategic stability

The impact of AI on strategic stability is conditional. It is based on three criteria: (*a*) the openness of the strategic stability environment, (*b*) instrumental rationalism in strategic stability thought and (*c*) the expansion of strategic stability factors.

The openness of the strategic stability environment

An important pathway for AI to have an impact on strategic stability among the great powers is the openness of the strategic stability environment. This condition depends on the overall international environment and is reflected in two aspects: changes in the distribution of power and the fragility of strategic stability.

[20] Luo, X. (罗曦), '美国构建全域制胜型战略威慑体系与中美战略稳定性' [US full-domain deterrence and its implications for Sino-US strategic stability], 外交评论 [Foreign Affairs Review], vol. 35, no. 170 (Mar. 2018), pp. 37–62.

[21] Treaty on the Non-Proliferation of Nuclear Weapons (Non-Proliferation Treaty, NPT), opened for signature 1 July 1968, entered into force 5 Mar. 1970.

[22] Li (note 18), p. 32.

Changes in the distribution of power

Openness of hegemony and great power status to incorporate more actors may stem from changes in the distribution of power among states. From the historical rise and fall of great powers, changes have been evident in their strength over time, such as the decline of ancient Rome and the British Empire. If the distribution of power among countries changes, emerging great powers will inevitably challenge the existing hegemonic order. The openness of great power status may be due to the loss of the dominant foundation on which the great powers have relied. For example, the advantages of the sea power era have been gradually surpassed and replaced by the convenience of land transport and air traffic.

The openness of great power status may also be due to the homogenization of technological superiority. Hegemonic powers gain advantage from innovation in fundamental production methods, distinguishing them from other countries.[23] However, this advantage will not last long. As technology spreads and other countries learn from those that have succeeded in the competition to survive in the international community, great powers will increasingly behave the same, the world will soon trend towards homogenization and hegemony will be weakened. For example, despite the efforts of the international community to control nuclear proliferation, the trend is for more states to acquire nuclear weapons. Due to the large temptation of nuclear capabilities, some countries are still eager to try to develop them.

The openness of great power status may also stem from the asymmetric effect of new forces. In the era of AI and cyber means, actors with weak conventional forces may use asymmetric approaches to provoke conflicts. Under the logic of cyberweapon and AI weapon asymmetry, strong powers would prefer defensive strategies, rather than launching attacks. This is because such countries are more dependent on high-technology networks and have higher anticipated losses in conflicts. Even if a weak country and a strong country show the same aggressiveness, an attack launched by a weak country should be more destructive. The inherent logic behind a weak country launching such an attack is to use the asymmetric effect to inflict greater damage.[24]

The fragility of strategic stability

Openness may also stem from the fragility of the strategic stability relationship among great powers. During the cold war, this fragility mainly arose from the balance of terror. Major nuclear powers believed that the use of nuclear force would lead to unacceptable retaliation, so they maintained a relationship of strategic stability primarily by ensuring the ability to use nuclear weapons to engage in counterattack. However, the current balance of nuclear terror has begun to be threatened, particularly following the withdrawal of the USA from the 1972 Anti-

[23] Liu, M. (刘鸣), '美国霸权实力何以能持久延续？' [How can US hegemonic power last forever?], 社会科学 [Journal of Social Sciences], vol. 29, no. 3 (Nov. 2007), pp. 43–53, p. 44.

[24] Liu, Y. (刘杨钺), '网络空间国际冲突与战略稳定性' [International conflict and strategic stability in cyberspace], 外交评论 [Foreign Affairs Review], vol. 33, no. 157 (Apr. 2016), pp. 106–29, p. 114.

Ballistic Missile Treaty (ABM Treaty) in 2002.[25] Unilateralism threatens strategic stability. With the destruction of this nuclear non-proliferation mechanism, the world has fallen into a multilateral nuclear security dilemma. However, mutual vulnerability in the nuclear field is not the only pillar that sustains strategic stability.

Currently, in addition to nuclear factors, strategic stability relations among the great powers are also characterized by interdependence. This encompasses increasing common interests, such as the joint response to international terrorism and the proliferation of nuclear weapons, as well as development of other advanced technologies, such as AI. Common challenges also include failed states, climate change and other threats that can jeopardize economic growth and prosperity.

With deepening economic and political interactions, the great powers find that interdependence on each other and the international system is constantly growing.[26] In this way, strategic stability relations among great powers can be maintained. This is not only because mutual vulnerability means that these states have the ability to cause unbearable damage to each other, but also because they need to achieve more important goals and to confront common challenges and threats. At the same time, while economic and political interdependence are among the cornerstones for the maintenance of strategic stability among great powers, events in the economic and political spheres may also induce instability. On the whole, common interests and interdependence contribute to the strategic stability of great powers, but this stability is fragile.

Instrumental rationalism in strategic stability thought

The second criterion in evaluating the role of AI in the strategic stability of great powers is based on the universal existence of instrumental rationalism in international relations. The realist thinking underlying instrumental rationalism believes in technology and power, typically emphasizing their use to directly achieve its purpose. Strategic stability at the highest level is the stability of will at the political level. However, under the prevailing role of instrumental rationalism, this cannot occur. Instrumental rationalism may create a dilemma, in that attention is often not paid to the effectiveness of the instrument. Instead, it is often dominated by an extreme panic about being overtaken by an adversary, thereby causing strategic instability.[27] The existence of instrumental rationalism in strategic stability thought has greatly enhanced the importance and emphasis

[25] 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.

[26] Finger, T. (托马斯•芬加) and Fan, J. (樊吉社), '中美关系中的战略稳定问题' [Strategic stability in Chinese–US relations], 外交评论 [Foreign Affairs Review], vol. 31, no. 138 (Jan. 2014), pp. 43–55, p. 44.

[27] Ge, T. (葛腾飞), '工具理性主义的困境与美国冷战决策模式的批判—<保罗•尼采:核时代美国国家安全战略的缔造者>评介' [The dilemma of instrumental rationalism and a critique of the US cold war decision-making model—a review of Paul Nietzsche: the founder of the US National Security Strategy in the nuclear age], 美国研究 [Chinese Journal of American Studies], no. 3, 2018, pp. 135–44, p. 139.

placed by great powers on AI among the most advanced technologies, thus enhancing its role in maintaining strategic stability among countries.

There are three reasons for the proliferation of instrumental rationalism.

The first is cold war mentality. Instrumental rationalism first arose from the fact that this cold war construct has not been overcome. In fact, strategic stability is a legacy of this manner of thinking. As just one example, the National Security Strategy of the US administration of President Donald J. Trump, issued in December 2017, positioned China as a strategic competitor.[28] A cold war mentality has caused great power competition to replace the terrorist threat as a new strategic concern for the USA. Trump believes that the world has entered a new era of competition, such that the military strength, economic strength and political competitiveness of a country are of paramount international importance. In January 2018, the US Department of Defense (DOD), in a summary of the US National Defence Strategy, unabashedly demonstrated that the USA wants to continue to use various means, including AI, to maintain its absolute military superiority and to ready itself for long-term strategic competition among major powers.[29]

Second, instrumental rationalism also stems from the lack of strategic mutual trust among great powers. It could be argued that the current comprehensive strategic stability among major powers must still be based on strategic stability in the traditional military field. While great powers, such as China and the USA, may have good intentions and are working hard to maintain their bilateral relations, ensuring a lack of conflict and confrontation among great powers cannot rely solely on the will and intent of the countries concerned. With the current widespread lack of mutual trust among major powers, their intentions are often difficult to clarify and almost impossible to verify.[30] The relative balance in military power is the key to ensuring that there is no conflict or confrontation. Therefore, instrumental rationalists believe that, even in times of peace, they must maintain stronger military power and strategic strength to ensure that potential attackers can be blocked at any time.

The third is fatalistic realism, from which instrumental rationality also derives. John Mearsheimer sums up the tendency for there to be conflict between a rising power and an established power as a tragedy of great power politics.[31] In Chinese–US relations, fatalistic realism maintains that China's rise will inevitably challenge the dominant position of the USA and will lead to the two countries fighting for hegemony. Belief in unavoidable conflict will inevitably shape each other's cognition and behaviour and poses one of the most serious threats to

[28] White House, *National Security Strategy of the United States of America* (White House: Washington, DC, Dec. 2017).

[29] US Department of Defense (DOD), *Summary of the 2018 National Defense Strategy of the United States of America: Sharpening the American Military's Competitive Edge* (DOD: Washington, DC, Jan. 2018). The full strategy is classified.

[30] Da, W. (达巍) and Zhang, Z. (张昭曦), '中美关系新阶段中的战略"失语"与战略稳定探索' [Strategic "aphasia" and strategic stability in a new stage of Chinese–US relations], 国际安全研究 [Journal of International Security Studies], no. 5, 2016, pp. 39–59, p. 57.

[31] Mearsheimer, J., *The Tragedy of Great Power Politics* (W. W. Norton & Co.: New York, 2014).

the strategic stability relationship between the two. Many Chinese experts instinctively regard any action taken by the USA that may have a negative impact on China as 'blocking' (遏阻) or 'containment' (围堵). Similarly, US scholars, media and politicians often claim that China's military modernization and activities around the world have a real but unspoken intention to challenge the dominant position of the USA.[32] If both sides believe that conflict is inevitable, the attitudes and policy actions of both countries will be affected. As a result, fatalistic realism may eventually erode all the pillars that maintain the strategic stability of great powers and result in a self-fulfilling prophecy.

The expansion of strategic stability factors

Another criterion for AI to influence strategic stability is through the expansion of strategic stability factors in the new era. Nuclear weapons are no longer the only consideration. To limit strategic stability to the field of strategic nuclear weapons does not guarantee comprehensive and effective security for a country. Nuclear weapons only defend a country's core security interests: ensuring that the country's central territory will not face a large-scale attack from foreign enemies. They will not provide effective support for a country's non-core interests.[33] For any great power, in addition to defending the core interests of the country's central territory, there are many other national interests. To effectively protect these interests requires a greater scope of stability that includes conventional military forces.

Moreover, the factors that influence strategic stability are not limited to the development of strategic military forces: they also cover new threats and instabilities. In other words, strategic stability has become an issue with multiple drivers. Factors such as unilateralism, nuclear proliferation, nuclear terrorism and the development of conventional weapons are evolving as new intervening variables that affect strategic stability.[34] Additionally, AI, cybersecurity, regional conflicts, energy issues, political and diplomatic influence, economic dependence, the level of scientific, technological and economic development, and the extent of participation in international affairs are all considerations for evaluating strategic stability among great powers.

The above-mentioned elements of strategic stability can be divided into three categories: technical factors, behavioural factors and institutional factors.[35] In other words, strategic stability is not only related to a country's deterrence under specific attack and defence patterns, but also to its behaviour and related mechanisms or systems. Technical factors establish the material basis for the comparison of strategic strength among countries. They not only determine the

[32] Finger and Fan (note 26), p. 48.

[33] Bo, E. (波尔特), '战略稳定概念对美国安全战略的影响及启示' [The impact and implications of the concept of strategic stability on US security strategy], 国际论坛 [International Forum], No. 5, 2016, p. 48.

[34] Li, D. (李德顺), '战略稳定性中的相互依赖因素' [The elements of interdependence in strategic stability], Doctoral thesis, Tsinghua University, May 2012, p. 19.

[35] Yu, Q. (俞倩倩), '从战略稳定性看反卫星武器的发展' [A look at the development of ASATs from the perspective of strategic stability], Master's thesis, Fudan University, 2008, pp. 17–18.

size of nuclear weapon forces, but also the level of military technology modernization and conventional forces. They are the fundamental factors in determining strategic stability. Behavioural factors are catalysts, guiding the ability to amplify or reduce material power. Institutional factors are the result of the behavioural interaction of states with one another. They can subtly change the actions of the state, establish a new norm of weapon technology development, and then reconstitute and shift technical and behavioural factors.[36]

As one of the most cutting-edge technologies in the technical factor category, AI plays an important role in all aspects of strategic stability. This is not only because it can affect traditional nuclear relations, conventional force comparison and so on, but also because it is a new variable with an impact on strategic stability. Following the cold war, conventional power advantage clearly shifted to the West. As a result, strategic stability guaranteed by the mutual deterrence of nuclear weapons became, in essence, the last pillar to maintain the balance of international military power. However, AI and cyber means offer an opportunity for a number of countries to garner an advantage. Therefore, as the maturity of AI increases, strategic stability is shaped by the extent of AI factors among technical elements.

IV. The ways in which AI could shape the future path of strategic stability

The core competencies of AI technology driven by deep learning algorithms include cognition, prediction, decision-making and integrated solutions.[37] Cognition refers to the perception and description of the world through the collection and interpretation of information, including such techniques as natural language processing, computer vision and audio processing. Prediction is based on obtaining a wide range of information, analysing different scenarios that may occur through multilayered neural networks, and predicting behaviours and outcomes that may occur in various scenarios in advance. Decision-making is comprised of effective analysis of collected information and completion of predictions regarding specific scenarios, to determine a course of action based on pre-set goals. Once AI is combined with other complementary technologies, it provides an integrated solution for extremely complex activities.

While the fundamental role of AI occurs via these four core competencies, the path of AI's impact on strategic stability can be subdivided into five aspects: (*a*) its empowerment effect on nuclear weapons, (*b*) its enhancement effect on conventional military forces, (*c*) its comprehensive penetrative effect on strategic capabilities, (*d*) its behavioural risk effect that leads to conflict escalation, and (*e*) its psychological anxiety effects.

[36] Li (note 34), p. 19.

[37] Feng, S. (封帅) and Zhou, Y. (周亦奇), '人工智能时代国家战略行为的模式变迁——走向数据与算法的竞争' [The pattern of change in national strategic behaviour in the age of artificial intelligence: towards competition between data and algorithms], 国际展望 [Global Review], no. 4, 2018, pp. 40–41.

Table 10.1. The empowerment effect of AI on nuclear weapons

AI application	Possible result	Impact on strategic stability	
Surveillance, target acquisition and reconnaissance	Higher or perceived higher risk of decapitating strike from an adversary by conventional weapons; higher mutual confidence due to increased transparency	✓ ✓	× ×
Early warning	Possible lower risk of accidental or misinformed launch of nuclear weapons	✓ ✓	×
Air and space defence and ballistic missile defence	Lower confidence in the survivability of second-strike retaliatory capability		×
Nuclear strike capabilities	Possible higher risk of accidental or unauthorized use of nuclear weapons; higher escalation risk		× ×
Command and control	AI as a trusted adviser; possible lower and higher risk, due to hacking or accidental or misinformed launch of nuclear weapons	✓	×
Protection systems for nuclear forces	Attack on nuclear forces or nuclear command and control by conventional weapon systems; higher risk or perceived risk of decapitating strike by an adversary		×

× = negative effect; ✓ = positive effect; AI = artificial intelligence.

Source: Derived from presentations by Nishida Michiru and Petr Topychkanov and subsequent discussion at the East Asia Workshop: The Impact of Machine Learning and Autonomy on Nuclear Risk, Beijing, 6–7 Sep. 2018.

The empowerment effect of AI on nuclear weapons

One of the ways in which AI plays a role in strategic stability is through its empowerment effect on nuclear weapons. Applications of AI that can empower nuclear weapons include in environmental detection, target location, early warning, air and space missile defence systems, nuclear weapon command systems, and protective systems for nuclear storage and transportation equipment.

Nearly all of the resulting scenarios may have an effect on nuclear strategic stability—positive or negative (see table 10.1). Nuclear experts and AI researchers seem to agree that advanced AI may seriously undermine the stability of nuclear strategy and increase the risk of nuclear war.[38] However, not all agree on how and why AI would have an impact. Indeed, AI has a double-edged impact on nuclear strategic stability.

The use of AI in two scenarios—in tracking missiles and as a decision aid on the use of nuclear weapons—illustrates the role that AI may play in nuclear warfare from both sides.[39] If AI is applied in tracking missiles, it will greatly improve the accuracy of monitoring potential enemy attacks. This increased transparency may enhance the strategic mutual trust between two parties, thereby reducing

[38] Geist, E. and Lohn, A. J., *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (RAND Corporation: Santa Monica, CA, 2018). This report is based on a series of 3 workshops in May–June 2017.

[39] Geist and Lohn (note 38).

the possibility of nuclear war and improving strategic stability. However, in the event of a crisis, using or providing AI-enabled intelligence, surveillance and reconnaissance (ISR) may also increase tensions and the possibility of unexpected escalation of the conflict. Moreover, if the AI missile-tracking function is flawed or hacked, the probability of triggering a nuclear war will greatly increase, thereby reducing strategic stability. Accordingly, AI decision-making support has a dual impact on nuclear war.

With the use of AI, the number of factors that have an impact in the fragile MAD-based nuclear balance will significantly increase. AI-enabled autonomy and sensor integration are of strategic importance since they can enhance ISR, automatic target recognition (ATR) and terminal guidance capabilities, which may seriously weaken nuclear force survivability. This would thereby shake a country's sense of security and undermine crisis stability. This has a greater impact on China and Russia, since they primarily rely on mobile intercontinental ballistic missiles (ICBMs) for deterrence. Of course, the ability to develop ATR, sensor integration and signal processing remains extremely difficult. According to one report, in an increasingly multipolar strategic environment, AI is likely to lead to the breakdown of the balance of nuclear weapons and the failure of current means of nuclear deterrence by 2040.[40]

Furthermore, involvement of AI technology will also introduce new variables into the stability of the global system of nuclear deterrence. In an era of weak AI, only a combination of AI technology and nuclear weapons can form an effective deterrent system. When AI technology is involved in all aspects of a nuclear deterrent structure, the original system of stability will change. As a data tool, AI provides countries with new offensive capabilities and has a direct impact on the reliability of nuclear weapon use. In a big data environment, however, there are also a number of subjective factors, such as the unpredictability of national will and strategic intent. When subjected to deep learning algorithms, intent may be clarified. These shifts could result in an imbalance in the MAD-based system of nuclear deterrence. The party with a command of AI technology will have the ability to clearly assess the possibility and destructiveness of the other party's nuclear counterattack, thus having more flexible strategic options, while the side with relatively backward technologies will possess less credible retaliatory capabilities.[41] The gap between military powers will again expand and countries' military strategic aims will accordingly readjust. In other words, the traditional international security system will become unstable.

The enhancement effect of AI on conventional military forces

The second path for AI to have an impact on strategic stability is through its upgrading of conventional military forces. While nuclear weapons were the most important pillar of strategic stability during the cold war, they were not an

[40] Geist and Lohn (note 38).

[41] Feng, S. (封帅), '人工智能时代的国际关系: 走向变革且不平等的世界' [International relations in the AI age: towards a world of change and inequality], 外交评论 [Foreign Affairs Review], no. 1, 2018, pp. 140–41.

Table 10.2. The enhancement effect of AI on conventional military forces

AI application	Possible result	Impact on strategic stability	
Target country and battlefield situational awareness	AI has the ability to collect battlefield information more comprehensively and efficiently. The use of natural language processing systems can more efficiently collect and process audio signals. Machine vision can enhance the ability of automatic weapon systems to identify and analyse battlefield conditions. This allows for increased transparency, strategic mutual trust, reduced motivation to launch war. However, false information may also increase risk perceptions.	✓ ✓	×
Military command human–machine cooperative decision-making	An intelligent command system with functions of reasoning, analysis, prediction, decision-making, etc., can greatly improve the accuracy and effectiveness of military command activities. AI can quickly process battlefield information and has the rapid response capability that humans lack. AI offers multithread processing capability, can simultaneously handle multiple military operations and can propose complex strategies that are beyond the capabilities of human thought.	✓	× ×
Assisting human activity	This includes portable electronic equipment and auxiliary power units to ensure that the soldiers get help in a variety of possible emergencies. This could strengthen existing power distribution among states and at the same time reduce the fear of activities of war.	✓	× ×
Collaborative operations (advanced manned or unmanned combat teaming)	This consists of using AI systems to coordinate actions, optimize operational strategies, and flexibly adjust to battlefield conditions and operational objectives to maximize battlefield advantage. At the same time, it will increase asymmetry.	✓	× ×
Network empowerment and autonomous high-speed weapons for cyberattacks and electronic warfare	This covers everything from real-time identification of defects and vulnerabilities by computer systems that completely lack human intervention, to the ability to quickly and automatically complete software repair and system defence in billions of lines of code, to creation of a hacker robot with both offensive and defensive capabilities. Because of non-lethality, use may increase. Developed and intermediate countries may be the biggest beneficiaries of empowerment with autonomous weapons.	✓	× ×
Lethal autonomous weapon systems	This features self-discovery of targets, self-determination and implementation of attacks. It is relatively controllable among rational state actors, but it is uncontrollable in the case of non-state actors such as terrorist organizations.		× ×

× = negative effect; ✓ = positive effect; AI = artificial intelligence.

operational option in the great power competition. This was due to the balance of nuclear terror and the consequences of mutual destruction. With the expansion of the concept of strategic stability, conventional military forces have become an important consideration. The world's military has transformed from an era of mechanization to one of information. Algorithm-based AI is an important promoter of this military revolution. It is expected to give birth to new combat styles and to change the mechanism of winning wars. In doing so, it has become an important means to change the rules of the game in warfare and to shape subversive military capabilities (see table 10.2).[42]

AI can also play a broad role in non-nuclear forces. For example, the proliferation of autonomous weapons is not limited to such traditional fields as UAVs, but rather may be fully rolled out in a variety of military fields. One scholar has warned that 'If autonomous weapons are developed and deployed, they will eventually find a home in every domain—air, space, sea, land, and cyber'.[43] Unlike previous technological changes, AI technology in the military field has led to changes in all aspects: from military weapons to strategic design and from global military power balance to military ethics, all will inevitably be affected.

In terms of environmental situational awareness on the battlefield, AI has the ability to collect more comprehensive battlefield information. For example, the use of machine vision can enhance the ability of the automatic weapon system to identify and analyse battlefield conditions. Moreover, the natural language processing system can efficiently collect and process audio signals. For the strategic environment of competitors in peacetime, AI is also able to employ big data for statistical analysis to sense changes in strategic posture in a timely manner. In terms of military command, an intelligent command system with functions of reasoning, analysis, prediction and decision-making, among other capabilities, can greatly improve the accuracy and effectiveness of military command formulation. Combat commanders are thereby able to grasp battlefield information and to gain more precise tactical advice.

In practice, before a conflict begins, the AI system would be able to provide a more comprehensive set of battlefield information, simulate the deployment and combat capabilities of both sides, complete a relatively accurate format of the battlefield from deductive simulations and quantify all potential outcomes from a range of probabilities derived from various military strategies. In line with this quantitative probability, an effective operational plan of force distribution and strategic deployment could be selected and carried out. This is because AI has two advantages that humans are unable to match. First, AI systems can exceed human capacity in quickly processing battlefield information and engaging in rapid response. Second, AI systems have multithreading processing capabilities that can undertake multiple military operations simultaneously and propose

[42] Long, K. (龙坤) and Zhu, Q. (朱启超), '"算法战争"的概念、特点与影响' [The concept, features and impact of 'algorithmic warfare'], 国防科技 [National Defense Science & Technology], vol. 38, no. 6 (2017), p. 39.

[43] Roff, H., 'To ban or regulate autonomous weapons—a US response', *Bulletin of the Atomic Scientists*, vol. 72, no. 2 (Mar. 2016), pp. 122–24; and Roff, H., 'Banning and regulating autonomous weapons', *Bulletin of the Atomic Scientists*, 24 Nov. 2015.

complex strategies that human thought patterns are unable to grasp.[44] AI can also help humans with complementary actions, such as portable electronic equipment and auxiliary power units, to help in a variety of possible emergencies.

Humans can coordinate operations with AI systems and optimize warfare tactics, while flexibly adjusting to battlefield conditions and combat objectives to maximize battlefield advantage. Automated technology allows a weapon system to achieve greater flexibility and self-determination to solve problems. An intelligent weapon system not only achieves a substantial separation between human and weapon, but also completely transforms the war activity into a task of the weapon system. This brings the casualty rate among combatants to near zero and maximizes the efficiency of weapon use and coordination among various weapon systems. More importantly, the use of intelligent weapons makes the traditional combat laws, such as killing enemy combatants, lose their practical significance.[45] At the same time, human–machine collaboration can also accomplish a good deal of the work that cannot be done by humans alone. The USA and Europe have made breakthroughs on a number of key technologies such as UAV synergistic flight, unmanned vessel bee colony combat, unmanned submersible network detection and manned or unmanned combat aircraft formation flight tests.

Beyond these capabilities, network empowerment and autonomous high-speed weapons for cyberattack and electronic warfare are areas in which AI is particularly promising. Cyberweapons must operate outside communication range and respond rapidly. As a result, attacks initiated and controlled by AI systems have great potential. Further, the non-lethal nature of cyberweapons may increase their use. This being said, development of autonomous cyberweapons may differ from traditional weapons in that requirements at the technical level are higher. In this case, it could be argued that a technologically developed medium-sized country may be the largest beneficiary of autonomous weapons and may rewrite conventional power distribution, thereby injecting more uncertainty and instability into the international system.[46]

Additionally, countries are also vigorously developing lethal autonomous weapon systems (LAWS) that can independently identify targets, make independent judgments and carry out attacks. These types of system have the ability to engage in automatic attack and may engage in inhumane killing. For a national actor with a rational decision-making model, such systems are relatively controllable. What truly affects strategic stability and the international system is the use of LAWS by non-state actors, such as irrational terrorist organizations. This is because the rapid pursuit of new advanced technologies has not only enabled great powers to develop and deploy new weapon systems for a revolution in military affairs, it has also provided new possibilities for the proliferation of weapons of mass destruction and LAWS.

[44] Feng (note 41), p. 140.

[45] Feng (note 41), p. 139.

[46] Work, R. O. and Brimley, S. *20YY: Preparing for War in the Robotic Age* (Center for a New American Security: Washington, DC, Jan. 2014), p. 33.

The enhancement effect of AI on conventional military forces also results in another two forms of change in strategic stability among great powers. Due to this upgrade, technologically advanced countries may encounter lower risks, combined with more effective attack tools, such that they are able to pose a serious challenge to their opponent's strategic deterrence. Thus, for countries that have historically had the ability to fend off an attack, the introduction of greater mobility, concealment and autonomy capabilities with the next generation of equipment may make their retaliation-based deterrence strategy ineffective.[47] The impact of AI technology will thereby aggravate the imbalance of conventional military power confrontation. Armed forces lacking AI technology will find it increasingly difficult to compensate for their disadvantages on the battlefield through tactics and strategies. Conventional confrontation will no longer be a rational strategic option and they will have to resort to asymmetric warfare.[48] At the same time, the development of new unmanned weapons may also change the traditional casualty counts of conflicts, thus increasing the rate of use of these weapons. These trends are undoubtedly not helpful for great power strategic stability. However, the strategic mutual trust generated by AI-enabled mutual battlefield situational awareness and attack capabilities will also increase, which will be beneficial to a certain extent.

The comprehensive penetrative effect of AI on strategic capabilities

The third way for AI to have an impact on strategic stability is through its full penetrative effect on strategic capabilities. From the vantage point of international politics, the most important value of AI lies in a potential shift in allocation of strategic capacity among countries.[49] Competition in science and technology is an important part of strategic jockeying among great powers and competition in the field of AI is a core element. Therefore, the speed and impact of promoting the application of AI in various fields will not only profoundly affect future victory in war, but also the strategic competitiveness of great powers (see table 10.3). In a broad sense, the strategic competitiveness of great powers is ultimately the foundation of strategic stability in peacetime.

The comprehensive penetrative effect of AI on strategic capabilities is mainly due to its high penetrative advantage. AI has become an irresistible technological trend and is entering all aspects of life and all social fields. From the perspective of technological development, the new generation of AI not only represents a new direction in science and technology but also has an extremely important impact on the path of research and development (R&D) tools, costs and even the paradigm of how R&D is conducted in other scientific fields. From an economic vantage point, a new generation of AI will reconstruct all aspects of economic activities,

[47] Liu, Y. (刘杨钺), '全球安全治理视域下的自主武器军备控制' [Arms control of autonomous weapons under global security governance], 国际安全研究 [Journal of International Security Studies], no. 2, 2018, pp. 49–71, p. 64.

[48] Feng (note 41), p. 140.

[49] Liu (note 47), p. 50.

Table 10.3. The comprehensive penetrative effect of AI on strategic stability

AI application	Possible result	Effect on strategic stability	Impact on strategic stability	
Economy	Leads to major changes in economic structure, promotion and upgrade of industrial transformation, and achievement of a new leap in productivity	Winner-takes-all (×) AI technology catch-up cycle shortened (✓)	✓	× ×
Society	Greatly improves the level of targeted public services and comprehensively improves the quality of people's lives		✓	× ×
Politics	Increases political governance of the country and enhances freedom of speech		✓	× ×
Security	Increases the maintenance of national security measures and enhances national competitiveness		✓	× ×

× = negative effect; ✓ = positive effect; AI = artificial intelligence.

such as production, distribution, exchange and consumption. It will also form new macro- and micro-level intelligent demands and promote the advancement of new technologies, new products and new industries. These major structural changes will promote industrial transformation and upgrade, achieving a new leap in productivity.

Within social development, a new generation of AI will bring new opportunities for social construction. The extensive application of AI in education, medical care, elderly care, environmental protection, urban operation and judicial services will greatly improve the level of targeted public service to comprehensively improve the quality of people's lives. In terms of global competition, AI has become a new focus of international competition. Major developed countries regard the development of AI as a major strategy to enhance national competitiveness and safeguard national security. Whoever takes the lead in achieving breakthroughs in the field of AI will dominate future development.

Further, AI is a strategic technology that affects a country's developmental destiny and is related to the comprehensive strength of the country. In order to seize the initiative within this technological competition, countries have made plans for national AI strategies. At present, the world's major technological powers—China, Russia and the USA among others—all attach great importance to AI development. A 2018 report that systematically examined the possible impact of AI on national security from an economic, information, and military perspective recommends that the USA pay special attention to controlling the potential catastrophic risk of AI being used by hostile countries or through unanticipated incidents.[50] A parallel

[50] Horowitz, M. C. et al., *Artificial Intelligence and International Security* (Center for a New American Security: Washington, DC, July 2018).

report suggests that the USA should introduce an overarching national AI strategy as soon as possible.[51] This would be to guarantee that the USA is the global leader in top-level design, overall planning and key investment in AI technology, to allow it to win strategic competition in the AI field against China, India, Russia and South Korea, among others.

The strengths of cutting-edge technologies, enabled by AI and big data, may contribute to the formation of a new strategic balance. On the one hand, this is because the fourth industrial revolution centred on AI may lead to a winner-takes-all situation among countries. The comprehensive penetrative effect of AI on strategic capabilities is not conducive to the strategic balance of major powers. This is an important reason why countries hope to seize the opportunity in this unstable state. On the other hand, the shortening of the AI technology cycle of catching up is favourable for the strategic balance among major powers. In the previous industrial revolutions, the time advantage of the leading country over those working to catch up was large. For example, when the United Kingdom launched the First Opium War in 1839, China was still an agricultural society. The technology gap between the two countries could have been measured in decades, if not centuries. However, in the era of intelligent revolution, developed countries have realized such achievements as smartphones, driverless cars and cashless payment. As a result, developing countries have the chance to make similar progress within a year or two of these advances. As a result, the time differential is becoming smaller and smaller. This is also conducive to the formation of a multipolar world and the improvement of strategic stability.

The behavioural risk effect of AI that leads to conflict escalation

The fourth pathway for AI's impact on strategic stability is through its shaping of behavioural risk that could contribute to conflict escalation. It can do this in three ways (see table 10.4).

First, AI may blur the boundaries between conventional and nuclear warfare, thereby causing conflict escalation. As Paul Bracken of Yale University, USA, points out, the continued improvement of technologies such as AI has the potential to weaken the strategy of minimum nuclear deterrence and to blur the boundaries between conventional and nuclear warfare.[52] AI technology can help achieve new breakthroughs in tracking, targeting and anti-submarine warfare or make it easier for high-precision conventional ammunition to destroy reinforced ICBM silos.[53] This ability to destabilize is particularly significant because policy-makers are more likely to threaten to use conventional weapons than to conduct any form of nuclear attack. In a crisis, the threat of conventional weapon use can put tremendous pressure on the opponent. Doing so may force the country to yield

[51] Horowitz, M. C. et al., *Strategic Competition in an Era of Artificial Intelligence* (Center for a New American Security: Washington, DC, July 2018).

[52] Bracken, P., 'The intersection of cyber and nuclear war', Strategy Bridge, 17 Jan. 2017.

[53] Holmes, J., 'Sea changes: the future of nuclear deterrence', *Bulletin of the Atomic Scientists*, vol. 72, no. 4 (July 2016), pp. 228–33.

Table 10.4. The behavioural risk effect of AI that leads to conflict escalation

AI application	Behavioural risk of AI application	Main impact	Impact on strategic stability
Blur the boundaries between conventional and nuclear war	The opponent believes that it is necessary to use nuclear weapons before being disarmed or to counter an attack that fails to engage in successful decapitation.	Causes conflict escalation	× ×
Increase armed behaviour options	AI applications such as autonomous weapons do not necessarily involve human casualties and can alleviate the pressure of domestic public opinion that a country may face when launching and participating in foreign military operations.		× ×
The intention behind using AI to perform tasks may be misunderstood	This may be interpreted as a serious provocation against a target country's security interests, leading to more stringent response measures. Hacking may lead to misjudgement or an escalatory response.		× ×

× = negative effect; ✓ = positive effect; AI = artificial intelligence.

but could also trigger a nuclear war. The reasons for conflict escalation are that the opponent believes that it is necessary to use nuclear weapons (*a*) before being disarmed, (*b*) to counter a partially successful attack, or (*c*) in the event of a crisis that leads to accidental use.

Second, AI may increase the options for armed behaviour and cause conflict escalation. For national actors, one of the advantages of AI applications such as autonomous weapons is that they do not necessarily involve human casualties (on the attacking side). They can alleviate the pressure of domestic public opinion that decision makers may face when launching and participating in foreign military operations, while increasing the tools available for performing tasks. In particular, AI applications such as autonomous weapons can reduce the potential cost of certain postures and activities that may be necessary but that could lead to an excessive deterioration of the situation.[54] At the same time, for the problems that could be solved through diplomatic negotiation and other means, the risk of conflict is increased.

Third, the intention behind the use of AI may be misunderstood, increasing the risk of conflict escalation. With the development of technology and the evolution of the global situation, national actors may increasingly use AI weapons. But how well they send unambiguous signals to demonstrate their intent is a challenge when performing these tasks. Instead, these activities may be interpreted as a serious provocation to security interests, leading to a more stringent response from the target country. This could result in unnecessary conflict escalation.

[54] Liu (note 47), p. 67.

Moreover, autonomous weapons are highly dependent on perception and exchange of information about the external environment. As a result, the likelihood of accidents and human-induced malicious interventions increases. For example, if a drone is subjected to a hack or other form of electromagnetic interference while performing a reconnaissance mission and this results in abnormal behaviour such as a crash, an impact or an explosion, the target may misjudge or make an escalatory response.[55]

The psychological anxiety effect of AI

The fifth way in which AI can affect strategic stability is through its psychological anxiety effect. This can lead to strategic mutual suspicion and arms racing and thus affect strategic stability (see table 10.5).

First, there is a concern among countries that their level of AI technology will be surpassed. Technology is the foundation of various strategic capabilities in nuclear, space and conventional forces. It is therefore generally believed that a new generation of AI will become an important strategic deterrent. As described above, AI may overturn the foundation of the nuclear deterrence strategy by 2040.[56] Just like the cold war of the 1940s and 1950s, each side has a reason to fear that its opponents could gain a technical advantage. In late 2017, President Putin hinted that AI may be the way in which Russia rebalances US power in defence.[57] Russian state media subsequently reported that AI is the key to Russia's defeat of the USA.[58]

Second, there are concerns among states that the AI-related rules system will be pre-emptively formulated by the major powers. Elements of strategic stability include technical and behavioural factors, as well as institutional ones. The rules system of AI technology and applications can rebuild technical and behavioural factors. At present, however, AI research is still in its infancy. As a result, the international norms at the relevant technical and behavioural level are still in essence unwritten. Historical development shows that the successful pioneers of technological development are often the makers of the rules and regulations. Generally, latecomers can only passively accept rules and regulations. Even if it is possible for them to formulate new rules, this is difficult. Therefore, major countries have stepped up their R&D related to AI, hoping to take the lead in this rule-making round of competition.

The third concern of states is how their loss of great power status could have an impact on their voice in international diplomacy. AI will become another status symbol of great power. Without occupying the commanding heights of AI, it will be difficult to have a prominent stake in the future international arena.

[55] Liu (note 47), p. 65.

[56] Geist and Lohn (note 38).

[57] President of Russia, 'Расширенное заседание коллегии Министерства обороны' [Extended meeting of the board of the Ministry of Defence], 22 Dec. 2017.

[58] '新的冷战？专家警告说，人工智能是全球军备竞赛的"首选武器"' [New cold war? Experts warn that artificial intelligence is the 'preferred weapon' of the global arms race], 网易号 [NetEase], 31 Jan. 2018.

Table 10.5. The psychological anxiety effect of AI

Psychological anxiety effect on strategic stability	Main impact	Impact on strategic stability
Concern that AI technology will be surpassed	Psychological anxiety leads to a blind pursuit of strategic advantage, rather than strategic stability Strategic mutual doubt can be caused by psychological anxiety	× ×
Concern over pre-emptive AI-related rulemaking		× ×
Concern over the loss of great power status and its impact on a country's international diplomatic voice		× ×

× = negative effect; AI = artificial intelligence.

Nuclear weapons were once the most important symbol of great power status. Today, the strategic capabilities of AI not only illustrate military power, but also demonstrate the level of a country's technological and industrial development. Having AI strategic capabilities will greatly enhance a country's voice within the international diplomatic struggle.

Of course, in addition to the concerns affecting strategic stability, great powers have also emphasized other concerns about AI and national security in important documents, such as national security strategies and science and technology development strategies. Among these are fear of losing control of AI technology. In essence, AI is easy to obtain through technical means, extremely difficult to control and has a low threshold for abuse. It can easily fall into the hands of extremist individuals, criminal gangs or even terrorist organizations, thus posing a major threat to political security and social stability. Another example is the fear of major security risks in AI applications. The application of AI technology has many uncertainties. As such, without predictive, early-warning and preventive capabilities, systematic and catastrophic risks in what could be called the 'AI era' are inevitable.

The impact of psychological anxiety caused by AI can be divided into two categories.

First, the anxiety brought on by the blind pursuit of strategic advantage is a destructive factor when it comes to strategic stability. Because of the instrumental rationality of strategic stability thinking, the strategic goal of a great power is often not strategic stability, but rather the pursuit of strategic advantage. Yet strategic stability is worth pursuing instead of strategic advantage. Strategic stability is a state in which great powers can pursue strategic advantage. According to a 2017 report, phenomena similar to the development of nuclear weapons by the USA and the USSR after World War II are taking place.[59] Countries may agree to propose a digital Geneva Convention that limits AI weapons, but this does not prevent independent nationalist groups, militias, criminal organizations, terrorists

[59] Allen, G. and Chan, T., *Artificial Intelligence and National Security* (Harvard Kennedy School, Belfer Center for Science and International Affairs: Cambridge, MA, July 2017).

and other countries from developing AI and carrying out AI attacks. Moreover, a country can withdraw from any treaty. So, it is almost certain that one party will turn AI into a weapon, even if this is just based on a desire to engage in self-defence. Between strategic advantage and strategic stability, the blind pursuit of AI-related strategic advantages is a potential hazard for the maintenance of strategic stability, because technology is viewed as an important factor in changing the balance of offence and defence. According to the theory of offence and defence, when the balance between the two shifts to make offence dominant, the weapon system with higher mobility and self-protection will enhance the attack advantage and increase the possibility that a pre-emptive attack will be launched.

Second, strategic mutual doubt caused by psychological anxiety is also a destructive factor of strategic stability. From the point of view of AI, no one can accurately predict what kind of conditions will be produced by unmanned vehicles and intelligent warfare. Lowered warfare thresholds, expanded arsenal scales and uncertain technological evolution paths make these AI-related arms races a new source of strategic mutual distrust among states.[60] Incomplete mastery of AI will only increase uncertainty about the ability of a particular opponent's AI to engage in attack and defence. According to some nuclear experts and AI researchers, China and Russia seem to believe that the USA is trying to use AI to threaten the viability of their strategic nuclear forces and trigger mutual suspicion.[61] As a result, disastrous consequences can occur in a crisis. Strategic mutual distrust has also led to a lack of information sharing among major powers in the field of AI. This exposes decision makers to the risk of potentially unwise judgments and reduces strategic stability.

V. Conclusions

Strategic stability did not end along with the cold war. On the contrary, the concept of strategic stability broadened following the conclusion of the cold war. Nuclear strategic stability during the cold war has developed into today's complex strategic stability. Its categories have expanded from nuclear power relations via military and security relations to overall strategic relations. Its protagonists have grown beyond the United States and the Soviet Union to include various global actors.

There is a feasible basis for AI as a 'second nuclear weapon' to have an impact on strategic stability. This is based on the openness of the strategic stability environment, which includes hegemony, great power status and the fragility of great power strategic stability relations. It is also based on instrumental rationalism derived from cold war thinking, fatalistic realism and low strategic trust among great powers. Most importantly, the numerical growth of these factors suggests that AI has great potential for strategic stability. Among the three elements of

[60] Liu (note 47), p. 63.
[61] Geist and Lohn (note 38).

strategic stability—technical factors, behavioural factors and institutional factors—technical factors establish the material basis for the comparison of strategic strength among countries. Technical factors not only determine the level of nuclear forces, but also the ability to engage in military modernization and the level of conventional armed forces. These are fundamental elements in determining strategic stability.

There are five pathways for AI to have an impact on strategic stability: its empowerment effect on nuclear weapons, its enhancement effect on conventional military forces, its comprehensive penetrative effect on strategic capabilities, the behavioural risk effects that lead to conflict escalation and the psychological anxiety effect. Although some factors can enhance strategic stability, the impact of AI may be negative in most cases, such as its blurring of the boundaries between conventional and nuclear wars, increasing the choices of armed behaviour and resulting in misunderstanding of intent when employed. The escalation of conflict, the psychological pursuit of strategic advantage instead of strategic stability, and strategic mutual distrust among countries are destructive factors that have an impact on strategic stability.

AI applications have great potential and may have a significant impact on strategic stability. However, many of the limitations of these applications also merit the attention of strategists. Among AI's many characteristics are its military and civilian use, easy proliferation and data dependence. It will bring significant challenges to existing laws, security and ethics. In terms of security, AI systems are inherently fragile and unpredictable. As such, system accidents and enemy cyberattacks can be catastrophic. Malicious actors may use these vulnerabilities to infiltrate nuclear weapon systems, while the injured state may be unaware. The 2018 US Nuclear Posture Review specifically addresses the impact of cyberthreats on nuclear command, control and communications (NC3) systems.[62]

In addition, the development of AI weapons represented by LAWS and arms racing may endanger human peace, stability and even survival. On the legal front, the rapid development of AI and militarization trends have seriously affected the core principles of distinction, proportionality and humanity in the existing international law of armed conflict. On the ethical side, the rise of machines brought about by AI has brought enormous challenges to traditional human–machine relations. Whether, what and how human moral standards should be embedded in increasingly intelligent machines needs to be studied in depth.

Given the potential impact of AI on strategic stability, it is necessary to design a framework for maintaining strategic stability in the AI era as soon as possible. Regarding technical factors, countries can cooperate on researching the vulnerability of AI systems, while maximizing the role of AI. On behavioural factors, major countries should not only establish a communication channel for crisis management but also consider a response plan for machine learning, judgement and execution. Nuclear attacks cannot be withdrawn, so the real dilemma is in how to prevent nuclear crises and how to mitigate the transformation of

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

traditional behaviour into nuclear crises. Once a potential nuclear crisis has occurred, it must be prevented from further escalation. In terms of institutional factors, major countries need to jointly build AI-related mechanisms to prevent the illegal proliferation and malicious use of AI technology, rationally regulate the military application of AI, and prevent excessive dependence on AI. Most importantly, countries should build strategic mutual trust in the era of AI on the basis of all these factors, thereby promoting strategic stability and advancing the process of world peace and development.

11. Regulatory frameworks for military artificial intelligence

VADIM KOZYULIN*

Among systems enabled by artificial intelligence (AI), autonomous systems represent threats that merit greater attention and analysis, with the ultimate goal of establishing international regulatory frameworks. This essay starts (in section I) by outlining in general the military threats posed by AI. It then (in section II) explores these various applications of AI in the military domain and (in section III) possible approaches to their regulation. The essay examines (in section IV) the role of strategic stability before concluding (in section V) by drawing these together into practical recommendations for future regulation.

I. Military threats from AI

Military threats posed by AI can be split into three main groups.

The first is the apprehension that lethal autonomous weapon systems (LAWS) will be able to select and engage their targets using unknown algorithms and without meaningful human control or direct human supervision. Numerous AI analysts, human rights organizations and lawyers believe that LAWS would violate international humanitarian law, moral norms and ethics and would lead to an accountability gap for unlawful acts committed by autonomous weapons.[1]

The second group is the potential undermining of strategic stability. Adoption of AI in new types of weapon—such as missile defence, cyberattack tools, electronic suppression, and hypersonic and space weapons—opens the door to military dominance or even a disarming first strike.

The third group of threats derives from the radical reduction in the length of time required for data analysis for command, control, communications, computers, intelligence, surveillance and reconnaissance (C4ISR). This represents a new field for deployment of AI-based computer programmes. In what may be called 'C4ISR outsourcing' or 'strategic time pressure', AI can lead militaries to gradually handing over data-collection, processing and operational functions to AI-based systems, thereby shortening reaction times and forcing overly rapid decision-making.

II. Military applications of AI

Overall, game-changing trends in military AI promise to progress from evolution to revolution in a step-by-step process. This will occur via three key stages. Under 'command by directive', the user must tell software agents exactly what to do at

* The views expressed in this essay are those of the author and do not necessarily reflect those of any organization to which he is affiliated.

every waypoint of the process to achieve the objective. With 'command by plan', the software is able to determine how best to reach those waypoints to avoid certain obstacles. Finally, 'command by intent' provides the system with only the objective, such as to patrol and secure an area. It does not require further instructions on how to reach waypoints or any other tasks required to pronounce an area secure.[2]

At each of these stages, AI-driven programmes manifest themselves in a wide range of military domains, in which they are destined to augment and to even replace humans. These include (*a*) analysis of intelligence, surveillance and reconnaissance (ISR) data from various sources, such as satellite images, radar data and social networks; (*b*) map compilation based on aerospace and unmanned aerial vehicle (UAV) surveillance; (*c*) detection, localization and classification of infrastructure, arms and weapons; (*d*) use of intelligent biometrics, such as human identification via face, gait and gestures; (*e*) speech recognition in a complex noise environment; (*f*) search for keywords in voice and digital signals; (*g*) early warning with analysis and selection of methods of radio signal suppression; (*h*) automation of decision-making with the use of lethal weapons; (*i*) homing in on targets under low visibility and jamming conditions; (*j*) information counteraction, such as fake news or 'virtual truth'; (*k*) assistance in decision-making and managing resources, such as control of troops and logistics; and (*l*) unmanned traffic control.

Such applications can be evaluated in practice through such programmes as the United Kingdom's autonomous last mile resupply, which enables proactive logistical support for troops in challenging situations using machine learning and mathematical modelling.[3] While in the context of a European power, it nonetheless could be applied in East Asia. Under this system, AI-based systems read the terrain and calculate the quickest and safest routes. In addition, UAVs and unmanned ground vehicles will become an important element of future logistics. New technologies will usher in new types of military unit, including for missile defence systems, cyber command, space forces, AI-based ISR, information warfare, early warning, electronic countermeasures, laser weapons, autonomous vehicles, unmanned underwater vehicles (UUVs), anti-UAV and hypersonic vehicles.[4] Based on these developments there will be new fusions of forces. For example, air and missile defence would be likely to demonstrate a twofold increase in effectiveness when integrated with electronic warfare systems.[5]

Using AI in the military sphere will result in the gradual introduction of robotics and automation in every possible sphere, from materials to logistics. The logistics of the future is capable of having a serious impact on strategic stability in every area, from the high automation of logistical processes to autonomous delivery of munitions on the battlefield. Information exchange among service branches

[2] Hennig, J., Schwartz, P. and Bailey, K., 'Mission command on semi-automatic', *Army AL&T*, Apr.–June 2017, pp. 51–55.

[3] Walker, A., 'Autonomous last mile resupply: TITAN robot put through its paces', Qinetiq, 9 Apr. 2018.

[4] On e.g. South Korea's Dronebot Jeontudan military unit see chapter 6 in this volume.

[5] Speed, J. and Stathopoulos, P., 'SEAD operations of the future: the necessity of jointness', *Journal of the Joint Air Power Competence Centre*, no. 26 (spring/summer 2018), pp. 38–43.

will develop both vertically and horizontally, from aircraft pilots in the air to platoon leaders on the ground. AI will filter information so that each party will only receive data that is useful to them, with data noise removed. This is the idea behind the DiamondShield integrated air and missile defence, which is currently under development by Lockheed Martin.[6] Data collected in air, in space and on land, including through Project Maven of the US Department of Defense (DOD), will be processed by neural networks and distributed in real time to commanding officers of all levels.[7] AI will direct the actions of military units, creating so-called algorithmic warfare.

AI is also destined to track clandestine action in times of peace. The Collection and Monitoring via Planning for Active Situational Scenarios (COMPASS) programme of the US Defense Advanced Research Projects Agency (DARPA) is one such example.[8] Its goal is to analyse behaviour in a grey zone situation—understood to be a limited conflict on the boundary between regular competition among states and what is traditionally deemed to be war. In such an environment, strategic time pressure leads to automation and outsourcing to AI-based command and analytical systems to conduct assessments of national threats and the use of weapons. The symbiosis of analytical and command programmes on the basis of neural networks increases the risk that the human–machine interaction model will leave little room for humans, who might have only limited time and options to approve decisions made by the machines.

These configurations of AI-based analytical and control systems are intended to be highly classified, thereby causing additional concerns among the public. At some point the human brain will not be able to keep pace with the intelligent supercomputer in controlling swarms of unmanned vehicles on land, in air and at sea in a changing environment. Under such conditions, the military sphere may witness something similar to the banking sector when robo-traders were introduced to the stock exchange—humans now only map out trading priorities to robo-traders, which act autonomously and make thousands of transactions a day.[9] A hypothetical future combat warfare control board may leave humans an option of AI-compiled scenarios that, once initiated, will be executed by AI-driven programmes. Humans may only have to choose the scenario and push the start button. As such, outsourcing of C4ISR functions to AI will lead to strategic time pressure and bring risks to global and regional security due to insufficient human control, lack of understanding between machine and human, and a lack of time for evaluation of situations and decision-making.

[6] Lockheed Martin, 'DiamondShield integrated air & missile defense', accessed 24 Apr. 2019.

[7] Pellerin, C., 'Project Maven to deploy computer algorithms to war zone by year's end', US Department of Defense, 21 July 2017.

[8] Barlos, F., 'Collection and Monitoring via Planning for Active Situational Scenarios (COMPASS)', US Defense Advanced Research Projects Agency (DARPA), accessed 24 Apr. 2019.

[9] Pearlstein, R., 'The robots-vs.-robots trading that has hijacked the stock market', *Washington Post*, 7 Feb. 2018. See also Scheftelowitsch, D., 'The state of artificial intelligence: An engineer's perspective on autonomous systems', ed. V. Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. I, *Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019), pp. 26–31, p. 29.

III. Possible regulatory approaches

In the light of these various programmes and potential future threats, there have been a variety of efforts to explore the issue of meaningful human control. This has been widely covered in a number of reports by the International Committee of the Red Cross (ICRC), the International Committee for Robot Arms Control (ICRAC), the United Nations Institute for Disarmament Research (UNIDIR) and SIPRI, among other think tanks and institutions. International experts include the following elements when exploring this problem: (*a*) accountability gaps; (*b*) risks of violation of human rights, the laws of war, human dignity and ethics; (*c*) threats to fundamental moral principles due to the decision to use force; and (*d*) concerns over the indiscriminate nature of these systems and lack of control.

Facing these challenges, in December 2016 the fifth review conference of the 1980 Convention on Certain Conventional Weapons (CCW Convention) set up a group of governmental experts (GGE) with the mandate to explore and agree on possible recommendations on options related to emerging technologies in the area of LAWS.[10] In 2018 the GGE adopted possible guiding principles, including that 'Accountability for developing, deploying and using any emerging weapons system in the framework of the CCW must be ensured in accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control.'[11] Within this structure, states have the discretion to choose to undertake measures on these weapon systems in accordance with their legal and security concerns and depending on global conditions. Among these measures, a political declaration may be the most easily achievable as it lays the groundwork for follow-up measures, such as verification instruments and legally binding obligations of the state parties.

Another potential avenue would be a politically binding agreement that would constitute a formal but non-binding document. As an alternate pathway, while the world has not yet gained enough experience in regulation of LAWS to implement good practice guidelines for their control, there are a few existing templates that could be adopted. Among these are the UK's doctrine on its approach to unmanned aircraft systems.[12] The US DOD directive on autonomy in weapon systems could also serve as a guide.[13] More broadly, a code of conduct would be appropriate as a means of codifying permissible and prohibited activities related to LAWS.

[10] 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, or 'Inhumane Weapons' Convention), opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983; 5th CCW Convention Review Conference, Final Document of the Fifth Review Conference, 23 Dec. 2016, decision 1; and UN Office at Geneva, '2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS)', accessed 24 Apr. 2019.

[11] Group of Governmental Experts of 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, 23 Oct. 2018, para. 21(c); and Kostopoulos, L., 'System characteristics and human involvement with lethal autonomous weapons systems (LAWS)', Medium, 23 Sep. 2018.

[12] British Chiefs of Staff, *Unmanned Aircraft Systems*, Joint Doctrine Publication no. 0-30.2 (Ministry of Defence, Development, Concepts and Doctrine Centre: Swindon, Aug. 2017).

[13] US Department of Defense, 'Autonomy in weapon systems', Directive no. 3000.09, 21 Nov. 2012, updated 8 May 2017.

In terms of internationally driven approaches that could be considered for adaptation in the East Asian context, the *Tallinn Manual on the International Law Applicable to Cyber Operations* would serve as another example of how to define and regulate disruptive technologies.[14] Apart from the political split caused by this document, its formal content offers a rigorous academic and legal approach for addressing issues inherent within LAWS.

Finally, there is the traditional means of addressing global threats from a weapon system: prohibition under the CCW Convention. However, implementation of a pre-emptive ban on the development, testing, transfer, deployment and use of LAWS would require complex organizational and technological solutions, which seem distant possibilities at the moment.

IV. The role of strategic stability

Strategic stability continues to play a pivotal role in reaching agreement on the regulatory frameworks described above. The 1990 Soviet–US joint statement on non-proliferation and strategic stability outlines the theoretical foundations of strategic stability, defined as a state of strategic relations between two powers in which neither has the incentive for a nuclear first strike.[15] While under negotiation, the parties raised two forms of strategic stability: crisis stability and arms race stability. Crisis stability was taken to mean a situation, even a crisis, in which neither party has serious opportunities or incentives to deliver the first nuclear strike. Arms race stability was driven by the presence of incentives to increase a country's strategic potential. In the years that followed, the guidelines for arms control enshrined in the 1990 joint statement came to be complemented by concepts of first-strike stability and even cross-domain strategic stability.[16]

Military AI has the potential to undermine stability within any of these conceptual variations as long as it increases first-strike capacity and provides a means to avoid retaliation. New technologies cause militaries to assume that robotic and unmanned weapons can evade layered missile defences, defend military bases against missile attacks, and serve as a replacement for nuclear weapons and precision-guided conventional weapons. Some high-ranking US strategists in the DOD have already stated that autonomous robots could ensure global military dominance. They believe that unmanned combat aerial vehicles (UCAVs) will replace nuclear weapons and high-precision munitions and will make it possible to implement the so-called Third Offset Strategy.[17] Obviously, such technologies

[14] Schmitt, M. N. (ed.), *Tallinn Manual on the International Law Applicable to Cyber Operations* (Cambridge University Press: Cambridge, 2013); and Schmitt, M. N. (ed.), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press: Cambridge, 2017).

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

[16] Kent, G. A. and Thaler, D. E., *First-Strike Stability: A Methodology for Evaluating Strategic Forces* (RAND Corporation: Santa Monica, CA, Aug. 1989); and Mallory, K., 'New Challenges in Cross-Domain Deterrence', Perspective, RAND Corporation, 2018. See also chapter 10 in this volume.

[17] Ellman, J., Samp, L. and Coll, G., *Assessing the Third Offset Strategy* (Center for Strategic and International Studies: Washington, DC, Mar. 2017). See also chapter 10 in this volume.

as machine learning and autonomy open new opportunities for using nuclear munitions for tactical missions and vice versa. Just one example would be a high-precision, reduced-capacity B61-12 nuclear bomb.

At the same time, an increasing number of strategic tasks can be handled using non-strategic weapons. For example, the development of hypersonic vehicles with high defence-penetration capabilities leads to a lower nuclear conflict threshold. The United States' X-37B orbital test vehicle, the XS-1 spaceplane and the X-43A experimental hypersonic vehicle will change the model of confrontations in space. By combining the Space Tracking and Surveillance System (STSS) with the Command and Control, Battle Management, and Communications (C2BMC) system, the USA has demonstrated entirely new strike capabilities of ballistic missiles.[18] The strategy of neutralizing missile systems at launch by using cyber and radio-electronic left-of-launch devices (that would defeat the threat of a nuclear ballistic missile before it is launched) opens up a new road map for missile defence. Quantum computing and automated hack-back cyberweapons, which identify hacking attacks on a system and try to identify their origin, are further ways in which software can be used offensively.

The rapid spread of UAV technologies around the world and the budding competition for the global market between major manufacturers of strike UAVs are causes for alarm. Today, the USA has tens of thousands of unmanned vehicles, including several hundred UCAVs.[19] Small UCAVs that in the future may deliver strikes as an autonomous swarm distributing functions without an operator's input are now under development and not simply in the USA. While China has not officially disclosed the number of UAVs in service with the People's Liberation Army (PLA), it is likely that it is roughly equal to that of the USA.[20] China both manufactures and actively exports strategic UAVs capable of both intelligence and strike missions.[21] Following the US MQ-25 Stingray programme, China is developing ship-based UAVs and unmanned vehicles capable of interacting with manned aircraft. In addition, the UK, Israel, Turkey, Iran and Japan are also among the world's leading UAV manufacturers. Military strategists of small and large states have come to believe that unmanned vehicles will form the future backbone of their air forces. As just one example, in 2015 the US secretary of the navy, Ray Mabus, said that the F-35 would probably be the last manned combat aircraft and that unmanned systems will become 'the new normal in ever-increasing areas'.[22]

[18] Missile Defense Advocacy Alliance, 'Command and Control, Battle Management and Communications (C2BMC)', accessed 10 Aug. 2019.

[19] Walker, J., 'Unmanned aerial vehicles (UAVs)—comparing the USA, Israel, and China', Emerj, 3 Feb. 2019.

[20] China Power, 'Is China at the forefront of drone technology?', Center for Strategic and International Studies, 13 Feb. 2019; and Easton, I. M. and Hsiao, L. C. R., *The Chinese People's Liberation Army's Unmanned Aerial Vehicle Project: Organizational Capacities and Operational Capabilities* (Project 2049 Institute: 11 Mar. 2013).

[21] Waldron, G., 'China finds its UAV export sweet spot', FlightGlobal, 14 June 2019.

[22] LaGrone, S., 'Mabus: F-35 will be "last manned strike fighter" the Navy, Marines "will ever buy or fly"', USNI News, 15 Apr. 2015.

V. Conclusions

Although the current international climate is unfavourable for the restoration of confidence and the reduction of tensions, many international measures for arms control and confidence building remain valid. As such, some continue to argue for revival of dialogue on security challenges. The Vienna Document 2011 on Confidence- and Security-Building Measures remains one of the key instruments of transparency that awaits its next modernization.[23] According to this document, countries should exchange data on major weapon and equipment systems, including the numbers of each type and information on plans for deployment in the zone of application (i.e. Europe, Central Asia and adjoining sea and airspace). Remotely operated or autonomous UCAVs could be included in Annex III of the Vienna Document, alongside combat aircraft and helicopters. Other transparency measures in this document that could also be relevant for LAWS include visits to airbases; demonstration of new types of major weapon and equipment system; prior notification of certain military activities; and observation of certain military activities.

When political conditions make it possible to negotiate an upgrade of the Vienna Document—or even a duplication in East Asia—the issue of autonomous ground and aerial vehicles could be included in the agenda. The 1990 Treaty on Conventional Armed Forces in Europe (CFE) established comprehensive limits on five key categories of conventional military equipment (battle tanks, armoured combat vehicles, artillery, combat aircraft and attack helicopters) in the area from the Atlantic Ocean to the Ural Mountains.[24] When changing political realities allow the start of new discussions on the control of conventional forces in Europe or in East Asia, the approaches that guided the state parties to conclusion of the CFE Treaty can be revisited. New categories of conventional weapon should supplement the previous five, with indication of new limits and transparency measures—this would revive traditional arms control in the areas of cruise and hypersonic missiles, UCAVs and autonomous ground combat vehicles.

Further, to facilitate these confidence-building measures, rapidly developing blockchain technology has the potential to open up new opportunities for cataloguing of and reliable technical control over the use of robots both during military exercises and in combat.[25] Globalization and cross-border projects, transnational corporations, international cooperation, observation satellites, radio-electronic intelligence and even social networks make the world more transparent. There are already many artificial ‘eyes and ears’ that could be useful in warning about new threats even before they materialize. Using these

[23] Vienna Document 2011 on Confidence- and Security-Building Measures (Vienna Document 2011), adopted 30 Nov. 2011, entered into force 1 Dec. 2011.

[24] Treaty on Conventional Armed Forces in Europe (CFE Treaty), signed 19 Nov. 1990, entered into force 9 Nov. 1992.

[25] On the use of blockchain in verification of arms control see also Kaspersen, A. and King, C., ‘Mitigating the challenges of nuclear risk while ensuring the benefits of technology’, ed. Boulanin (note 9), pp. 119–27, p. 126.

more positive manifestations of AI transparency and control can supplement the development of a regulatory framework for addressing its military applications for the sake of enhancing international security.

12. The environmental impact of nuclear-powered autonomous weapons

HWANG IL-SOON AND KIM JI-SUN*

The development of nuclear-powered, unmanned nuclear weapon systems—in particular unmanned underwater vehicles (UUVs)—threatens to irrevocably harm the 1968 Non-Proliferation Treaty (NPT) and the globe.[1] This is due to the unpredictability of these autonomous platforms and the potential for extensive fallout of long-living radioisotopes into the biosphere caused by the explosion of the nuclear warheads and nuclear reactors mounted on these weapon systems. For large-scale targets, a potential blast from these weapon systems could release a significant quantity of transuranic radionuclides, contaminating the biosphere for tens of thousands of years. Their development encourages non-nuclear weapon states and terrorists to consider their own radiation-dispersal devices, undermines the peaceful use of nuclear energy and hinders longer-term prospects for nuclear security cooperation. Thus, the development of these nuclear-propelled autonomous nuclear weapon systems should be banned.

To illustrate the damage that could be caused by nuclear-powered autonomous weapons, this essay first reviews the current Russian development of two such weapons (in section I). It then (in section II) estimates the extent of the damage that use of such weapons risks and concludes (in section III) by recommending some short- and long-term steps to remediate the risks of the deployment and proliferation of such platforms.

I. Development of nuclear-powered autonomous weapons

Russia's development of an ultra-long-range UUV mounted with a nuclear warhead was first leaked by Russian media in November 2015.[2] Poseidon (also known as Status-6) has been characterized as an unmanned torpedo with a length of 24 metres and diameter of 1.6 metres.[3] In March 2018 Russian President Vladimir Putin formally announced successful development of this nuclear-powered unmanned nuclear weapon system in his annual address to the Russian Federal Assembly.[4] He also unveiled an intercontinental nuclear cruise missile being developed for deployment in the next 5–10 years. According to his account

[1] Treaty on the Non-Proliferation of Nuclear Weapons (Non-Proliferation Treaty, NPT), opened for signature 1 July 1968, entered into force 5 Mar. 1970, IAEA INFCIRC/140, 22 Apr. 1970.

[3] 'Status-6/Kanyon—Ocean Multipurpose System', GlobalSecurity.org, accessed 10 Aug. 2019; Mathew, A., 'Russian MoD releases footage of Poseidon (Kanyon/Status-6) nuclear unmanned underwater vehicle', DefPost, 19 July 2018; and Gallacher, S., 'Russia launches sub that will carry doomsday nuke drone torpedo', ArsTechnica, 25 Apr. 2019.

[4] President of Russia, 'Presidential address to the Federal Assembly', 1 Mar. 2018.

* The views expressed in this essay are those of the authors and do not necessarily reflect those of any organizations to which they are affiliated.

and later analyses, these weapon systems have the ability to evade existing missile defence systems by taking advantage of autonomy and high-speed nuclear propulsion technology.

Up to six stealth Poseidon units can be loaded and launched from a submarine, each with a range of 10 000 kilometres, cruising at a maximum depth of 1000 metres at a speed of 185 km per hour.[5] As of December 2018, media reports detailed its sea trials, with video imagery that was consistent with this description.[6] Up to six nuclear warheads in the forward section are shielded from a small modular reactor that delivers propulsion power by using an integrated steam turbine system.[7] The small modular reactor vessel has about the same diameter as the nuclear warhead located in the forward combat compartment. Both warhead and reactor are housed in Poseidon's monolithic shell, with no provision for separation before the final target attack.

Poseidon's targets are said to include enemy naval bases and military ports.[8] The nuclear warheads range up to 2 megatons, less than the 100 megatons claimed in 2015.[9] In addition, this weapon system could be salted, using cobalt-60, a radiation source whose radiotoxicity can last for several decades. According to press reports, detonation would probably occur underwater near a coast, leading to a powerful blast followed by enormous tsunami effects.[10] Therefore, it is anticipated that the reactor would explode, along with the nuclear warhead, resulting in significant environmental impact.

The intercontinental cruise missile announced by Putin at the same time is the Burevestnik, which reportedly has a nuclear propulsion engine, giving it unlimited range.[11] The missile can reportedly cruise at low altitude and high speed to defy enemy missile defence systems. It was initially expected that only a static ground test of the missile's nuclear engine would be conducted. However, according to unnamed officials of the Department of Defense (DOD), the cruise missile crashed during testing in the Arctic in late 2017 and early 2018.[12] While there was no indication of radioactive fallout after these incidents, this platform may have been involved in another alleged accident in August 2019 that triggered a radioactive blast.[13] Similar to Poseidon, both Burevestnik's nuclear warhead and

[5] 'Status-6/Kanyon—Ocean Multipurpose System' (note 3).

[6] 'Poseidon underwater drone trials confirm its speed, unlimited range—source', TASS, 6 Feb. 2019.

[7] 'Status-6/Kanyon—Ocean Multipurpose System' (note 3).

[8] 'Russia's nuclear underwater drone could trigger 300-foot tsunamis, headed for battlefield by 2027', South Front, 21 May 2018; and 'Source: Russian Poseidon underwater drone capable of carrying 2 megatonne nuclear warhead', TASS, 17 May 2018.

[9] 'Russia begins testing of "Poseidon" underwater nuclear drone', PressTV, 26 Dec. 2018.

[10] 'Russia's nuclear underwater drone could trigger 300-foot tsunamis, headed for battlefield by 2027' (note 8).

[11] Stratfor Worldview, 'Russia's New Weapons: From Doomsday Nuclear Torpedoes to Skyfall Missiles', *National Interest*, 20 Aug. 2019.

[12] Macias, A., 'Putin claimed a new nuclear-powered missile had unlimited range—but it flew only 22 miles in its most successful test yet', CNBC, 21 May 2018.

[13] 'Russia's new arms give the US room for pause', Stratfor Worldview, 16 Aug. 2019; Kramer, A. E., 'Russia confirms radioactive materials were involved in deadly blast', *New York Times*, 10 Aug. 2019; and Reuters, 'US-based experts suspect Russia blast involved nuclear-powered missile', *Moscow Times*, 10 Aug. 2019.

its integral propulsion reactor would explode at the target. This raises serious concerns about future radioactive fallout.

II. Radioactive contamination from explosion of propulsion reactors

Propulsion reactors produce slow nuclear reactions during long-range travel. Thus, the nuclear radioactivity from the explosion of a propulsion reactor is generally longer-lived than that from a nuclear warhead. Although the power and operation periods of the nuclear reactors of the Russian unmanned vehicles are not known, a rudimentary assessment can be made using standard values. A typical 300-kiloton W87 thermonuclear missile warhead releases about 1300 terajoules of energy.[14] A typical 10-megawatt electric micro-modular reactor emits approximately the same amount of energy when continuously operated for one year.

When calculating environmental costs, most of the energy released originates from the fission reaction in proportion to the quantity of fissile products. However, remaining radioactivity is dominated by transuranic actinides that have unusually long persistence. For the same amount of energy released, their radioactivity—including the optional salting radioactive sources—is expected to be comparable to fissile products. In contrast, a short-lived nuclear warhead explosion would not produce a significant quantity of long-living actinides. As a consequence, the radioactivity of nuclear power reactors lasts about 1000 times longer than that of nuclear weapons, extending their impact over 100 000 years.[15] This is due to the transuranic by-products, which possess extremely long half-lives.

Considering the magnitude and length of fallout from radioactivity, all nuclear-powered warhead delivery systems should be considered as massive radiological dispersal devices, contaminating large areas for an extended period of time. In fact, in 1948 the United Nations Commission for Conventional Armaments included 'radioactive material weapons' in its definition of weapons of mass destruction.[16] At the 2012 Nuclear Security Summit in Seoul, participants agreed to make concerted effort to control radioactive source materials to stem the production of radiological dispersal devices.[17] If some countries are allowed to deploy nuclear-powered dirty bombs, it will compel others to undertake countermeasures. It is possible that radioisotope sources or spent nuclear fuels from research or power generation could be diverted in response to the deployment of nuclear-powered dirty bombs.

[14] Medalia, J., *'Dirty Bombs': Technical Background, Attack Prevention and Response, Issues for Congress*, Congressional Research Service (CRS) Report for Congress (US Congress, CRS: Washington, DC, 24 June 2011).

[15] American Physical Society (APS), 'Nuclear energy fission', APS Physics, accessed 22 Aug. 2019.

[16] UN General Assembly Resolution 36/97 B, 'Conclusion of an international convention prohibiting the development, production, stockpiling and use of radiological weapons', A/RES/36/97, 9 Dec. 1981; and United Nations, Security Council, Commission for Conventional Armaments, Resolutions adopted by the Commission at its thirteenth meeting, 12 August 1948, and a second progress report of the Commission, S/C.3/32/Rev.1, 18 Aug. 1948, p. 2.

[17] 2012 Seoul Nuclear Security Summit, Communiqué, 26–27 Mar. 2012.

Beyond their payloads, the autonomy and stealth of these nuclear-powered delivery platforms make them difficult for the attacker to control and for the target to intercept once they are launched from a base or submarine.[18] Nonetheless, there are some deficiencies in these vehicles that could be exploited to mitigate their impact. The extended range of nuclear-powered vehicles could allow time for defenders to take defensive actions that employ machine learning, including altering boundary conditions and poisoning signal data. Moreover, the reliance of artificial intelligence (AI) on empirical data-pattern matching is certain to leave unforeseen loopholes that could be used to undermine the effectiveness of these platforms. Yet, as much as these gaps can be exploited to mitigate the threat, such weapons may still provoke doomsday reactions by all sides. This level of escalation is unacceptable.

III. Conclusions

Ongoing development of nuclear-powered autonomous nuclear weapon vehicles may trigger a nuclear war due to the risks in both the unpredictability of autonomous decision-making and the predictability of long-lived radiation fallout within the biosphere. This is a categorical step away from arms control and directly contravenes the NPT, in particular Article VI under which each of the parties 'undertakes to pursue negotiations in good faith on effective measures relating to cessation of the nuclear arms race at an early date and to nuclear disarmament, and on a treaty on general and complete disarmament under strict and effective international control'.[19] Thus, the development, deployment and use of these weapon systems must be raised in disarmament meetings and addressed with a resolution under the NPT, combined with international consensus on banning any autonomy that takes the human out of the loop. While building this international consensus, the application of AI in weapon systems—in particular, nuclear weapon systems—should be avoided.

Furthermore, with the advent of nuclear propulsion, the extended periods of operation of unmanned systems will inevitably involve equipment breakdowns due to ageing associated with vibration, impacts, corrosion and wear. Unlike manned vehicles, for which there are provisions and procedures to inspect, mitigate and repair early degradation, emerging defects in these unmanned systems may deteriorate at unacceptably faster rates. Subsequent accidents with pressure boundary failures can result in significant radioactive contamination of seawater. To address such deficiencies, the 1986 Convention on Early Notification of a Nuclear Accident establishes a system of notification of nuclear accidents 'from which a release of radioactive material occurs or is likely to occur and which has resulted or may result in an international transboundary release that could be

[18] Etzioni, A. and Etzioni, O., 'Pros and cons of autonomous weapons systems', *Military Review*, May–June 2017, pp. 72–81.

[19] Treaty on the Non-Proliferation of Nuclear Weapons (note 1), Article VI.

of radiological safety significance for another State'.[20] This convention requires the responsible state to report the time, location and nature of the accident and other data essential for assessing the situation. Reporting is mandatory for any nuclear reactor no matter where it is located and even when such facilities are used for power generation in space objects.[21] Considering the high vulnerability of nuclear-powered unmanned weapon systems to serious accidents, the application of the Convention on Early Notification of a Nuclear Accident should be enforced as soon as possible.

[20] Convention on Early Notification of a Nuclear Accident, opened for signature 26 Sep. 1986, entered into force 27 Oct. 1986, IAEA INFCIRC/335, 18 Nov. 1986, Article 1(1).

[21] Convention on Early Notification of a Nuclear Accident (note 20), Article 1(2)(a)

13. East Asian security dynamics as shaped by machine learning and autonomy

ARIE KOICHI*

The current nature of nuclear weapon dynamics is highly complex and spans everything from nuclear warheads and delivery vehicles to command and control. This complex network is further integrated with advanced conventional weapons, such as missile defence. While related platforms are currently operated and managed by humans, they are expected to include some form of machine learning and autonomy in the future, as artificial intelligence (AI) enters nearly every field of human activity. Given this ever-growing complexity, it is essential for the peace and security of East Asia that strategic stability among three nuclear weapon states—China, Russia and the United States—is maintained. Within this trilateral set of relations, they must also cope with the Democratic People's Republic of Korea (DPRK, or North Korea), a state that continues to develop nuclear weapons. As they proceed with AI applications within their own nuclear weapon systems and related conventional systems, the risk of nuclear escalation may increase.

This essay considers how the security dynamics of East Asia will be shaped by machine learning and autonomy. It first (in section I) examines the ways in which AI can be applied in nuclear forces. It then (in section II) discusses how these applications could have an impact on nuclear deterrence and arms control in East Asia. It concludes (in section III) by outlining steps to avoid nuclear escalation and maintain strategic stability in East Asia.

I. Applications of AI in nuclear forces

There are four types of possible nuclear force-related application of AI: (*a*) in nuclear weapons; (*b*) in enhanced intelligence, surveillance and reconnaissance (ISR) against enemy nuclear forces; (*c*) in nuclear command, control and communications (NC3); and (*d*) in conventional weapon systems that are relevant to nuclear forces.

First, AI-based nuclear weapons, if deployed, could evade existing countermeasures to successfully strike a target. For example, Russia is now developing the Poseidon nuclear-capable, unmanned underwater vehicle (UUV), which the Russian Ministry of Defence considers to be invulnerable to countermeasures.[1] Nonetheless, the countries affected by this development are likely to devise new means of countering these platforms, with even a new AI-based nuclear weapon

[1] Gady, F.-S., 'Russia begins sea-trials of nuclear-capable "Poseidon" underwater drone', The Diplomat, 21 July 2018. On Poseidon see also chapter 12 in this volume.

* The views expressed in this essay are those of the author and do not necessarily reflect those of any organization to which he is affiliated.

potentially developed to counter this UUV. Such a development would lead to a destabilizing tit-for-tat nuclear arms race.

Second, AI enables ISR systems to detect, identify and track an adversary's nuclear missiles before launch, even if these platforms are mobile and can be hidden in tunnels, forests and caves.[2] Project Maven, which the US Department of Defence initiated in 2017 with the help of Google and other technology companies, includes the use of AI to identify objects from video data gathered by unmanned aerial vehicles (UAVs). The project is reportedly exploring ways to use AI to find enemy nuclear missiles.[3] These types of application would make retaliatory nuclear forces vulnerable and undermine deterrence stability.

Third, introduction of machine learning into NC3 appears to be the most contentious type of application. On the one hand, AI, with increasingly super-human performance, would be better in supporting nuclear decision makers than human advisers when it comes to nuclear deployment and use. On the other hand, a machine learning- and autonomy-oriented NC3 system may also be compromised by hacking, poisoning of training data and manipulating of inputs.[4]

Fourth, introduction of AI in conventional weapon systems could also affect nuclear deterrence. Russia has repeatedly expressed concern that the US Conventional Prompt Global Strike (CPGS) programme, when combined with its global missile defence capabilities, could negate Russia's retaliatory nuclear forces.[5] China has also shown similar concerns about a US non-nuclear strike, whether against its conventional or nuclear forces.[6] If AI were introduced into the CPGS programme, this would seriously undermine Chinese–Russian–US trilateral nuclear stability.

II. The impact on nuclear deterrence and arms control in East Asia

Any type of application of AI in nuclear forces has the potential to destabilize the strategic environment.[7] In East Asia, nuclear stability among China, Russia and the USA is essential to overall regional security. However, each might have an incentive to strike first during a severe crisis in the region if AI were to increase the vulnerability of its retaliatory nuclear forces. Another source of nuclear risk

[2] See e.g. chapters 2 and 11 in this volume.

[3] Stewart, P., 'Deep in the Pentagon, a secret AI program to find hidden nuclear missiles', Reuters, 5 June 2018.

[4] Geist, E. and Lohn, A. J., *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (RAND Corporation: Santa Monica, CA, 2018). See also Avin, S. and Amadae, S. M., 'Autonomy and machine learning at the interface of nuclear weapons, computers and people', ed. V. Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. I, *Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019), pp. 105–18.

[5] Acton, J. M., 'Conventional Prompt Global Strike and Russia's nuclear forces', Carnegie Endowment for International Peace, 4 Oct. 2013.

[6] Roberts, B., 'Strategic stability under Obama and Trump', *Survival*, vol. 59, no. 4 (Aug.–Sep. 2017), pp. 47–84, p. 51.

[7] Trenin, D., 'Mapping global strategic stability in the twenty-first century', Russian International Affairs Council, 1 Nov. 2018.

is North Korea, which relies on a modest number of nuclear ballistic missiles for deterrence. However, the survivability of those missiles has been significantly undermined by rapid technological advances in remote sensing, data processing and communication.[8] ISR systems that have been enhanced by AI advance these trends. In extreme circumstances, North Korea might be tempted to use nuclear missiles for fear of losing them.

Applications of AI in nuclear forces have the overall potential to contribute to inadvertent or deliberate escalation to the nuclear level. With the introduction of AI in NC3, an AI adviser could be susceptible to error and hacking, leading to an accidental nuclear escalation in East Asia. The Soviet nuclear false alarm incident in 1983, commonly called the Petrov incident, partly illustrates this point.[9] Even if it worked properly, an AI-enabled system might deliberately advise human operators or decision makers to use nuclear weapons in order to prevail in a conventional military standoff between nuclear-armed states in the region.

US extended nuclear deterrence for Japan and the Republic of Korea (South Korea) would also be undermined if the above-mentioned nuclear risks materialized along with increased applications of AI in nuclear forces. These two countries may well be keenly interested in how US extended nuclear deterrence would work in the context of emerging AI applications in the nuclear forces of China, Russia and the USA. As a result, it is essential that these non-nuclear-armed states be consulted in addressing such risks.

In addition to risks, there are some arms control benefits to these AI-enabled systems. In East Asia, AI-enhanced ISR systems could increase transparency among nuclear-armed states. These systems could be employed for treaty verification and compliance monitoring for nuclear forces.[10] To enable these uses, the information acquired must be shared among the nuclear-armed states in the region. Also, each of these countries must be willing to disclose data on its nuclear forces. This would be more feasible when accompanied by traditional confidence-building measures (CBMs) such as de-alerting or de-targeting of nuclear forces. In other words, a greater degree of trust is needed among the states concerned.

However, current trends suggest that nuclear-armed states in East Asia are headed in the opposite direction. There is a risk that they will compete with each other to explore better applications of AI in their nuclear forces. As noted above, Russia is developing the nuclear-capable Poseidon UUV.[11] China is also reportedly working on introducing an AI adviser in its nuclear submarines in order to support the situational awareness and decision-making of submarine commanders.[12] Unless necessary arms control measures are put into place, these trends could

[8] 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, pp. 37–46.

[9] Tucker, P., 'Risk of "accidental" nuclear war growing, UN research group says', Defence One, 19 Apr. 2017. See also Topychkanov, P., 'Autonomy in Russian nuclear forces', ed. Boulanin (note 4), pp. 69–75, p. 70.

[10] Geist and Lohn (note 4), p. 6.

[11] See also chapters 7 and 12 in this volume.

[12] Chen, S., 'China's plan to use artificial intelligence to boost the thinking skills of nuclear submarine commanders', *South China Morning Post*, 4 Feb. 2018.

play out in the form of an arms race over nuclear force-related applications of AI in East Asia.

These trends also apply to conventional weapon systems that are relevant to nuclear forces. A conventional missile defence system installed with AI could intercept North Korean nuclear ballistic missiles more efficiently. However, such a system could also raise concerns in China and Russia that their retaliatory nuclear missiles could be intercepted. Some CBMs must be undertaken among China, Russia and the USA to address introduction of AI in missile defence systems. Otherwise, nuclear arms race stability would be jeopardized as escalatory military countermeasures are taken to protect their nuclear forces.

III. Conclusions

In the discussion of the four types of application of AI in nuclear forces, specific uses—which include such platforms as Poseidon and AI-enhanced nuclear submarines—remain largely focused on the operational or tactical level, rather than the strategic. There is a danger that use of AI capabilities at the operational level could be detached from strategic considerations.[13] If this is the case, nuclear use triggered by AI at the tactical level could lead to a further nuclear escalation. As a result, a cascading series of nuclear risks could jeopardize strategic stability in East Asia.

In order to minimize potential impacts of AI on nuclear risks in the region, plausible future scenarios regarding those risks must be explored and analysed for policy formulation. Scenario building should include experts from all the stakeholders in regional security. In this way, not only nuclear-armed states but also Japan and South Korea can contribute to future nuclear stability and hopefully derive a solution as to how AI could be better integrated and applied to minimize nuclear risks and to promote transparency for nuclear arms control in East Asia.

[13] Karlin, M., 'The implications of artificial intelligence for national security strategy', Brookings Institution, 1 Nov. 2018.

14. Arms control and developments in machine learning and autonomy

NISHIDA MICHIRU*

Arms control can be defined as 'all forms of military cooperation between potential enemies in the interest of reducing the likelihood of war, its scope and violence if it occurs, and the political and economic costs of being prepared for it'.[1] Confidence-building measures (CBMs) can thus be considered as one kind of arms control. CBMs can be categorized as taking a weapon-focused approach or a behaviour-focused approach. An example of a weapon-focused approach is to prohibit or control the amount of a certain type of weapon. A behaviour-focused approach would, for example, constrain a certain type of activity in relation to weapon development or use.

This essay identifies potential arms control measures, including CBMs, that could be applied to autonomous weapon systems and their impact on nuclear forces, drawing from previous practices in other areas of arms control. To better examine the future of the two approaches—weapon-focused and behaviour-focused—in relation to emerging technologies, it first provides a brief history of arms control measures (in section I). It then considers how future arms control approaches can be adapted to take into account the unique features of autonomous weapon systems (in section II). It concludes (in section III) by identifying the most promising of these measures.

I. A brief history of arms control

Traditional arms control measures, including CBMs, in the field of nuclear weapons can be divided into three relatively distinct eras. The first, the era before the Strategic Arms Limitation Talks (SALT) between the Soviet Union and the United States, comprised a largely behaviour-focused approach. This was typified by the 1963 Soviet–US Memorandum of Understanding Regarding the Establishment of a Direct Communications Link, known as the hotline agreement.[2] It was followed by the 1971 Soviet–US Agreement on Measures to Reduce the Risk of Outbreak of Nuclear War, which contained missile pre-launch notification measures.[3] The second era was that of SALT and the Strategic Arms Reductions Treaty (START) negotiations, which comprised a primarily weapon-focused approach. These led

[2] Memorandum of Understanding Regarding the Establishment of a Direct Communications Link, signed and entered into force 20 June 1963, *United Nations Treaty Series*, vol. 472 (1963), pp. 163–69. See also Davenport, K., 'Hotline agreements', Arms Control Association, Apr. 2018.

[3] Soviet–US Agreement on Measures to Reduce the Risk of Outbreak of Nuclear War, signed and entered into force 30 Sep. 1971, *United Nations Treaty Series*, vol. 807 (1972), pp. 57–62.

* The views expressed in this essay are those of the author and do not necessarily reflect those of any organization to which he is affiliated.

to the 1972 Soviet–US Anti-Ballistic Missile (ABM) Treaty and the 1991 Soviet–US Treaty on the Reduction and Limitation of Strategic Offensive Arms (START I), among others.[4] The third era has been the post-cold war era, which has included both weapon-focused and behaviour-focused approaches. These have included the Presidential Nuclear Initiatives (PNIs) of 1991, the 1996 Comprehensive Nuclear-Test-Ban Treaty (CTBT) and the 2010 Russian–US Treaty on Measures for the Further Reduction and Limitation of Strategic Offensive Arms (New START).[5]

For arms control to be successful, the following four factors are generally necessary. First, there should be significant strategic stability merit in such an agreement. Second, a clear definition of the object to be controlled is imperative: the object targeted by arms control needs to be clearly distinguishable from other non-controlled weapons. Third, the control measure needs to be verifiable. Even if an object is distinguishable and definable, if the control measure is not verifiable, then arms control is not possible. Fourth, the dual-use nature—that is, whether it has both military and non-military uses or both nuclear and non-nuclear uses—of an object of arms control should be minimal. For example, nuclear arms control started out with controlling missiles and launchers because they were more easily verifiable than nuclear warheads, which had to wait until New START offered specific on-site inspection procedures for their verification.

While arms control of nuclear weapons could be said to be mainly based on a weapon-focused approach, many arms control measures in the field of conventional weapons adopted a behaviour-focused approach in the form of CBMs. For example, the 1981 Comprehensive Study of the United Nations Group of Governmental Experts on Confidence-building Measures listed a number of military-related CBMs, including (*a*) exchange of information and communication on military activities; (*b*) reduction of military expenditure; (*c*) prior notification of military activities; (*d*) exchanges and visits; (*e*) establishment of a consultation mechanism; (*f*) measures to ease military tensions; (*g*) limitations or exclusion of certain military activities including demilitarized zones; (*h*) verification of CBMs, arms control and disarmament agreements; and (*i*) crisis management and settlement of disputes.[6] Many of these CBMs have been adopted in nuclear arms control agreements in the form of consultations, verification, hotline agreements and missile pre-launch notification.

[4] 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 Reduction and Limitation of Strategic Offensive Arms (START I), signed 31 July 1991, entered into force 5 Dec. 1994, expired 5 Dec. 2009.

[5] Kimball, D., 'The Presidential Nuclear Initiatives (PNIs) on tactical nuclear weapons at a glance', Arms Control Association, July 2017; Comprehensive Nuclear-Test-Ban Treaty (CTBT), opened for signature 24 Sep. 1996, not in force; UN Office for Disarmament Affairs, 'Comprehensive Nuclear-Test-Ban Treaty (CTBT)', accessed 25 Apr. 2019; Russian–US Treaty on Measures for the Further Reduction and Limitation of Strategic Offensive Arms (New START, Prague Treaty), signed 8 Apr. 2010, entered into force 5 Feb. 2011; and US Department of State, Bureau of Arms Control, Verification and Compliance, 'New START', accessed 25 Apr. 2019.

[6] United Nations, General Assembly, Comprehensive Study of the Group of Governmental Experts on Confidence-building Measures, A/36/474, 6 Oct. 1981, para. 128.

II. Arms control of autonomous weapon systems

A critical and fundamental difference between arms control of traditional nuclear and conventional platforms and that of autonomous weapon systems is that machine learning is highly dual-use. As a result, it is extremely difficult to differentiate between what will be permitted and what will be prohibited under any arms control agreement. Another critical difference is that machine learning is highly prone to proliferation. The international nuclear architecture has been largely driven by the five nuclear weapon states under the 1968 Non-Proliferation Treaty (NPT).[7] However, since machine learning proliferates quickly, the countries in possession of autonomous weapon systems are highly likely to exceed these five states.

These differences lead to the question of how arms control can best be applied to the future impact of autonomous weapon systems on nuclear forces. The type of autonomous weapon system related to nuclear forces that is easiest to imagine would be nuclear forces enabled by artificial intelligence (AI). However, autonomous weapon systems and AI are far from static. Instead, they are multidimensional and encompass a broad spectrum from offence to defence (see figure 14.1).

Along this spectrum, AI-enabled nuclear forces are the most offensive in nature. This is particularly the case with automatic target recognition (ATR) and autonomous targeting and engagement, which could result in a higher risk of escalation and accidental or unauthorized use of nuclear weapons. The second most offensive item on the spectrum is an attack on nuclear forces or their command and control by conventional autonomous weapon systems. This would entail greater, or at least perceived to be greater, risk of a decapitating strike by an adversary. The third most offensive case on the spectrum is AI-enabled intelligence, surveillance and reconnaissance (ISR). When pitted against an adversary's nuclear forces and combined with non-nuclear strategic weapons, such as the US Conventional Prompt Global Strike (CPGS), this heightens an adversary's concerns over a decapitating first strike.

The fourth case, AI-enabled missile defence, is where the spectrum starts to become more defence-oriented. This more capable missile defence contributes to lowering an adversary's confidence in the survivability of its second-strike retaliatory capability. This is followed on the spectrum by AI-enabled early-warning systems that have the potential to lessen the chances of an accidental or misinformed launch of nuclear weapons. The most defence-oriented case on the spectrum is AI as a trusted adviser, where an AI system's suggestions are treated similar to or better than those of human advisers in nuclear decision-making.[8] On one hand, this could reduce the risk of accidental or misinformed launches. On the

[7] Treaty on the Non-Proliferation of Nuclear Weapons (Non-Proliferation Treaty, NPT), opened for signature 1 July 1968, entered into force 5 Mar. 1970, IAEA INFCIRC/140, 22 Apr. 1970. The NPT defines a nuclear weapon state to be a state that manufactured and exploded a nuclear weapon or other nuclear explosive device prior to 1 Jan. 1967. According to this definition, there are 5 nuclear weapon states: China, France, Russia, the UK and the USA.

[8] Geist, E. and Lohn, A. J., *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (RAND Corporation: Santa Monica, CA, 2018), pp. 18–20.

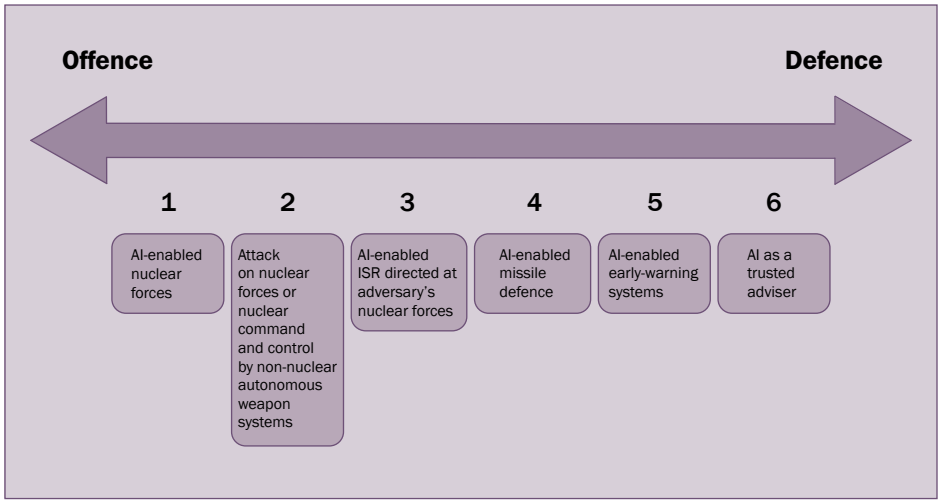

Figure 14.1. Spectrum of autonomous weapon systems in relation to nuclear forces
AI = artificial intelligence; ISR = intelligence, surveillance and reconnaissance.

other hand, the vulnerability in this last case relates to the susceptibility of such systems to hacking and other forms of interference that could lead to accidental or misinformed launch of nuclear weapons.

This spectrum is by no means exhaustive. However, it could serve as a basis for formulating standards, controls or CBMs in the field of autonomy relating to nuclear forces. Furthermore, in studying potential arms control and CBMs in the field of autonomous weapon systems relating to nuclear forces, two basic notions should be taken into consideration.

First, since the most important objective of any form of arms control is to ensure a higher level of strategic stability, any measure should seek to maximize stabilizing factors and minimize destabilizing factors. In other words, the emphasis should be shifted from the offensive end to the more defensive end of the spectrum.

Second, both weapon- and behaviour-focused approaches should be applied when possible to each of the conditions on the spectrum. For AI-enabled nuclear forces, the potential risk of accidental launch of nuclear weapons should be reduced or avoided. For this to occur, the principle of meaningful human control or meaningful human judgement could be instituted in an agreement—perhaps in a non-legally binding instrument—among nuclear-armed states. Nonetheless, it remains unclear as to how the implementation of such a principle could be effectively verified. Regular and mutual familiarization visits would be something to consider in this regard.

As for an attack on nuclear forces or nuclear command and control by non-nuclear autonomous weapon systems, it is necessary to mitigate a vicious cycle of distrust that could unintentionally lead to conflict. As such, it has been suggested that a verifiable agreement—not to base or to deploy autonomous weapon systems that might be used for a disarming or decapitating strike within a certain distance

of each other's mobile missile launchers—could be an option for consideration.[9] AI-enabled ISR directed at an adversary's nuclear forces should not be interpreted as a destabilizing factor, but rather as a stabilizing one in terms of transparency. However, when AI-enabled ISR is coupled with conventional or non-conventional disarming strike capabilities, such as CPGS, it rapidly becomes destabilizing. Since it is nearly impossible to prohibit the development and deployment of CPGS, the best option for weapon-focused control would be quantitative control, not prohibition.

AI-enabled missile defence is in essence a defensive system. However, as the history of missile defence shows, an adversary may perceive this enhanced missile defence coupled with offensive nuclear capabilities as a threat to the survivability of its nuclear forces.[10] This is a classic offence–defence dynamic with no real effective solution in the foreseeable future. Some kind of CBM—such as familiarization visits—may be the only way forward. Further along the spectrum, AI-enabled early-warning systems could be stabilizing factors and thus need to be enhanced. One way forward would be 'radical transparency' in which an AI algorithm that is used to provide decision support about escalation could be shared with adversaries.[11] Finally, AI as a trusted adviser could also be a stabilizing factor and should, therefore, be enhanced. As such, it could be worth exploring the feasibility of an agreement on the principle not to engage in physical attack or hacking.

On this spectrum (in figure 14.1), the weapon-focused approach could theoretically be applied to the first and possibly the fourth cases. However, traditional quantitative control would not solve the potential risks posed by these offensive manifestations of AI applications. The problem is not about numbers, but rather about two scenarios: the potential or perceived risk of (*a*) an accidental launch of nuclear autonomous weapon systems and (*b*) a decapitating strike by conventional autonomous weapon systems. In confronting these two scenarios, arms control with a weapon-focused approach could theoretically prohibit AI-enabled nuclear forces or autonomous weapon systems. However, the issue remains as to whether or not such an agreement would pass the test of distinguishability and verifiability. For example, if the software could simply be replaced before and after any verification, how could any form of arms control be sustainable and effective?

In contrast, the behaviour-focused approach could be applied to essentially any of the cases along the spectrum. As discussed above, this includes the concept of prohibition against autonomous weapon systems used to conduct a decapitating strike against nuclear forces within a certain distance. Prior notification of military activities—as with prior notification on ballistic missiles—has traditionally served as a key behaviour-focused measure. However, there are questions about whether this could function as originally intended in terms of autonomy as it relates to

[9] Geist and Lohn (note 8), pp. 16–18.

[10] Arbatov, A. and Dworkin, V. (eds.), *Missile Defense: Confrontation and Cooperation* (Carnegie Moscow Center: Moscow, 2013); and Saalman, L., 'China's evolution on ballistic missile defense', Carnegie Endowment for International Peace, 23 Aug. 2012.

[11] Geist and Lohn (note 8), pp. 21–22.

nuclear forces. For example, how would the autonomous weapon system of the adversary react to such a pre-notification? Could the adversary's autonomous weapon system differentiate between peacetime and crisis? Is there a way to limit escalation risk, due to unpredictable interaction between autonomous weapon systems?

Similarly, it is not clear if a crisis-management measure, such as a hotline, would be meaningful. In a world of autonomous weapon systems that have an impact on nuclear forces, there would be limited time to respond. There may not be any time for political leaders to exchange views with each other, discuss internally with their own staff and decide how to respond. Familiarization visits and verification would have a similar problem if the software could be easily replaced before and after such visits and verification. For the purpose of verification, this problem may be serious, but perhaps not so serious for the purpose of familiarization visits, which do not require as high a level of confidence as verification. Through regular and persistent visits, confidence could be gradually created over the long term. This also applies to consultation. However, this is a lengthy process and autonomy integration is advancing quickly.

III. Conclusions

Some arms control measures, including CBMs, for autonomous weapon systems related to nuclear forces could include such measures as publication or exchange of national policies and practices. This could be conducted with a view to accumulating international best practices in the future. There could also be implementation of regular or ad hoc mutual familiarization visits and consultations at bilateral, regional and international levels. Other areas of controls could include prohibition of AI-enabled nuclear forces or at least the adoption of the principle of meaningful human control for AI-enabled nuclear forces. There could also be exploration of a prohibition against attack on nuclear command and control by autonomous weapon systems or a ban against deployment of autonomous weapon systems that engage in a decapitating strike against nuclear forces within a certain distance.

In addition to controls placed on more offensive advances, there could be a greater focus on the defensive end of the spectrum, as with encouragement of AI-enabled early-warning systems and 'radical transparency'. Moreover, moderate use of AI as a trusted adviser could be a step forward, particularly if combined with a prohibition on attacking or hacking of these systems. As noted at the start of this essay, many of these arms control measures and CBMs are still notional and require a degree of systemic change among the potential signatories to any such agreement. Nonetheless, they can serve as a foundation for further exploration of concrete measures to address the proliferation of autonomous weapon systems that have an impact on nuclear forces and postures and on East Asian regional stability.

Conclusions

15. The impact of artificial intelligence on nuclear asymmetry and signalling in East Asia

LORA SAALMAN

This edited volume is the second 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 risk and strategic stability. It assembles the views of 13 experts from East Asia, Russia and the United States, 10 of whom participated in a workshop on the topic organized by SIPRI and the China Institutes of Contemporary International Relations (CICIR) in September 2018 in Beijing. This concluding chapter explores the role of asymmetry and signalling in East Asia as applied to AI and nuclear risk. It begins (in section I) with an overview of the risks and dynamics of the use of machine learning and autonomy by nuclear-armed East Asian states. It continues (in section II) by considering confidence-building measures (CBMs) that may be applied to the military applications of AI. It concludes (in section III) with a discussion of the means by which misalignment in assumptions and capabilities may be addressed.

I. Risks and dynamics of machine learning and autonomy

As this volume demonstrates, many of the risks and dynamics that play out in East Asian analyses and scenarios continue to centre on China, Russia and the USA. Yet, findings from the workshop and this volume indicate that every state in the region could be seen as wielding the power of a nuclear-armed state, regardless of whether or not it is an actual nuclear power, given the impact of machine learning and autonomy on strategic stability and nuclear risk. In fact, some of the more comprehensive discussions of arms control and CBMs are in the essays of experts from the non-nuclear-armed states of Japan and the Republic of Korea (South Korea), which face some of the most intractable risks associated with the integration of these technologies into regional nuclear dynamics.[1]

These two countries continue to play central roles in extended deterrence and to stand at the forefront of missile defence and nuclear escalation related not only to the Democratic People's Republic of Korea (DPRK, or North Korea), but also to Chinese–Russian–US trilateral dynamics. As such, they are uniquely qualified to provide third-party insights into how best to address technological and strategic change in East Asia. Moreover, when it comes to asymmetry, North Korea arguably has the most pervasive concern over its strategic inferiority and the potential for decapitation, increasing its perceived need for greater AI integration and rapid response.[2]

[1] See chapters 5, 12, 13 and 14 in this volume.
[2] See chapters 5 and 6 in this volume.

Nonetheless, when it comes to the asymmetries and signalling deficiencies that may have the most deleterious impact on nuclear risk, Chinese, Russian and US dynamics continue to dominate current East Asian analyses. In part, this is due to their long-standing tensions and the impact of their sizeable nuclear arsenals and AI advances on nuclear dynamics. Facing these asymmetric challenges, essays throughout this volume reveal that automation and autonomy are of strong interest for weaker nuclear-armed states with less capable early-warning systems and with smaller and less capable nuclear and conventional arsenals. To supplement these deficiencies, AI-enhanced platforms are seen as being capable of faster anticipation, discrimination, reaction and response. Yet, as much as this AI technology is viewed as a means of off-setting imbalances for weaker powers, it is also seen as being likely to tip the scales in favour of stronger powers.

This leads to a traditional security dilemma. As one Chinese expert at the East Asia workshop noted, automation of nuclear response is a 'fact, rather than a guess', as launch-on-warning is no longer simply being considered by Russia and the USA. Beyond enhanced launch capacity, both Chinese and Russian experts refer to the threat of US missile defence and Conventional Prompt Global Strike (CPGS) to the resiliency and survivability of their two countries' conventional and nuclear forces.[3] Further, Chinese workshop participants noted particular concern over US manned and unmanned manoeuvres around the first and second island chains that surround China to monitor, challenge and even potentially collide with Chinese submarines and vessels.[4] When such US platforms are enhanced by AI, these concerns are only likely to grow. As a result, East Asian regional actors are increasingly compelled to develop and deploy platforms with longer endurance and range—such as unmanned underwater vehicles (UUVs), unmanned aerial vehicles (UAVs) and spaceplanes to enhance both surveillance and nuclear deterrence throughout the region.

Facing such a future, the East Asia workshop and this volume suggest that countries facing both perceived and real asymmetries may be inclined to accept the risks of machine learning and autonomy to avoid the greater dangers of being caught off guard. While one side's dependence on these AI-enabled platforms may lead to some vulnerabilities that can be exploited, these enhanced systems are credited with strengthening the ability of countries to anticipate and to respond to threats. This time-compression issue when combined with the black box nature of machine learning in decision-support systems results in a dual dilemma of both attack and response time.[5] Facing this dilemma of response time, a Japanese expert maintained at the East Asia workshop that machine learning capabilities still largely remain at the 'observe' segment of the observe–orient–decide–act (OODA) loop. However, given that a number of new nuclear systems will be developed and fielded over the next 10 years, it is critical to have a means of understanding which nascent technologies are being considered for integration and where they are on

[3] See chapters 4, 11 and 14 in this volume.

[4] Huang, E., 'China's master PLAN: how Beijing wants to break free of the "island chains"', *National Interest*, 19 May 2017.

[5] See chapter 10.

the OODA loop. In particular, the biases and assumptions programmed into the algorithms are just as important as where these algorithms are inserted within the decision-making cycle.

Thus, in looking for answers on how these algorithms are being applied, the submarine-launched low-yield nuclear ballistic and cruise missiles mentioned in the 2018 US Nuclear Posture Review, Russian sea trials of the Poseidon nuclear-propelled and nuclear-armed UUV, and Chinese development of the DF-ZF hypersonic glide vehicle may serve as strong starting points and decisive tests for both AI integration and CBMs.[6] Rather than independent developments, these platforms constitute a complex intertwined security dilemma based on both national assumptions and technological advances. US discussion of low-yield platforms is widely perceived to be a response to Russia's alleged posture of escalating a conflict in order to de-escalate it, while both China and Russia are commonly thought to be seeking to maintain their second-strike capabilities with hypersonic vehicles and UUVs in the face of US pursuit of missile defence and CPGS.[7]

Thus, Chinese responses have in some cases started to mirror Russian countering of US capabilities to avoid being caught off guard. China's technological pursuits indicate that it is adopting a more offensive, forward-leaning stance in everything from launch-on-warning to development of prompt high-precision platforms. As noted by one Chinese expert at the East Asia workshop, while China has not changed its official stance of no first use, if these AI-enabled advances in US conventional and nuclear forces turn China's second-strike capability into a 'third strike'—such that the latter may be the victim of successive attacks and lack a chance for retaliation—China would have to undertake countermeasures in advance. Thus, while China's and Russia's shifts in nuclear posture remain under debate and are often denied, the fact remains that the USA is shaping its nuclear deterrent on assumptions and findings based on China's and Russia's technological advances, rather than simply their stated doctrines.

So while there are certainly differences in national responses within East Asia, there are some commonalities when it comes to the impact of machine learning and autonomy on nuclear risk. Most of these relate to the stability–instability paradox of these technologies.[8] In terms of transparency, East Asian experts tend to agree that machine learning improves reconnaissance, with the potentially destabilizing outcome that it will become more difficult to hide nuclear forces.[9] Because of the shallow depths of the South China Sea and the relative noise of their propulsion, Chinese experts argue that China's nuclear submarines already suffer

[6] US Department of Defense (DOD), *Nuclear Posture Review* (DOD: Washington, DC, Feb. 2018), pp. 54–55. See also chapters 4, 8 and 12 in this volume.

[7] See chapters 8 and 10 in this volume. See also Schneider, M. B., 'Escalate to de-escalate', *Proceedings* (US Naval Institute), vol. 143, no. 2 (Feb. 2017); and Oliker, O. and Baklitskiy, A., 'The Nuclear Posture Review and Russian "de-escalation": a dangerous solution to a nonexistent problem', War on the Rocks, 20 Feb. 2018.

[8] See chapter 10 in this volume.

[9] See chapters 5, 6, 8, 10, 11 and 14 in this volume.

from diminished survivability, which is only exacerbated by unmanned vessels.[10] This may be contrasted with the potentially stabilizing role of machine learning in concealing nuclear forces to enhance survivability and mutual vulnerability, through enabling anticipation or confusion of monitoring by satellites, UAVs and UUVs.[11]

In terms of timing, both Russian and US participants at the East Asia workshop cited the danger that automation bias—complacency or over-reliance on automated or autonomous systems—may occur that could lead to strategic time pressure and rash actions.[12] Yet, they also maintained that the very ability of the machine to process information faster increases the time available to humans to analyse and verify large volumes of data and to take decisions in a crisis. On communication, East Asia workshop participants expressed the concern that instability could arise from infiltration of command-and-control systems, since AI-based information gathering is vast and dispersed.[13] Japanese, South Korean and US experts argued at the various SIPRI workshops that this can result in disinformation and ambiguous signalling.[14] Nevertheless, they also noted that a greater number of channels of information flow, facilitated by machine learning, can also allow for a more comprehensive and balanced understanding of the threat environment.

II. Confidence building and the military use of AI

Through understanding the differences and commonalities in the views of experts from East Asia, particularly in terms of the stability–instability paradox, the East Asia workshop and this volume yield a series of potential CBMs. In terms of crisis management, some experts maintain that since machine learning is not yet state of the art, the world has a brief window in which to return to arms control.[15] However, this suggestion also elicited criticism at the various SIPRI workshops due to the persistent difficulty in defining what to control in lethal autonomous weapon systems (LAWS) and the ossified nature of current arms control structures, which are already saturated with long-standing disputes.[16] US experts further noted at the workshops that traditional arms control agreements were based on limits to development and testing, which in the case of AI is no longer feasible due to the speed at which it is being acquired and the difficulties of verification. Nevertheless, the Japanese experts in this volume suggest that a

[10] Wu, R., 'Survivability of China's sea-based nuclear forces', *Science & Global Security*, vol. 19, no. 2 (2011), pp. 91–120; and Zhao, T., *Tides of Change: China's Nuclear Ballistic Missile Submarine and Strategic Stability* (Carnegie Endowment for International Peace: Washington, DC, 2018).

[11] See e.g. chapter 2 in this volume.

[12] See also e.g. chapter 10 in this volume; and Horowitz, M. C., 'Artificial intelligence and nuclear stability', ed. V. Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. I, *Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019), pp. 79–83.

[13] See also chapters 5 and 6 in this volume.

[14] See also e.g. chapters 10 and 11 in this volume.

[15] See chapters 11 and 14 in this volume.

[16] See also e.g. chapter 11 in this volume.

focus on behaviour-based controls and scenario-building may overcome some of these barriers.[17]

On permissions and disengagement, preprogrammed algorithms or kill switches have been cited for inclusion in autonomous weapons for many years without much progress. Thus, technology experts at the East Asia workshop suggested that a requirement be built into the system's architecture, such that each segment of the system would be unable to exceed its own level of authority. Nonetheless, throughout such discussions, the difficulty of applying technological controls continues to confront the biases and assumptions that are preprogrammed—often unwittingly—into algorithms in each country. In confronting this reality, one US participant in the East Asia workshop argued that it is difficult for a military commander to rely on a system when there is little to no information on the principles programmed into its behaviour.[18] Another US participant further lamented how existing scenarios and training in table-top exercises continue to follow an overly linear and sequenced series of events, when an actual crisis is much more complex, with coterminous mingling of incidents and stakeholders.

Facing these challenges, East Asian participants at the various SIPRI workshops and contributors to this volume have suggested some areas in which progress could be made. Experts from China and Russia suggest a need to adapt strategic stability to accommodate the stabilizing and destabilizing nature of advances in machine learning and autonomy.[19] In doing so, there is a persistent and pervasive call to return to engagement of stakeholders, particularly in East Asia, where countries often view themselves to be at an asymmetrical disadvantage in AI, nuclear and conventional technologies. In this volume and at the East Asia workshop, a variety of forums have been suggested in which countries can be engaged even though their incentives for AI integration differ.

At the bilateral level, table-top exercises and strategic dialogues between Russia and the USA and between China and the USA were suggested. Recognizing that these dyadic relationships are deteriorating, however, Russian, Chinese and US experts recommended using multilateral groupings such as the United Nations Disarmament Commission as a deliberative and universal body that is open to new debate, as well as using a new group of governmental experts (GGE) established by the UN General Assembly as a platform for in-depth technical discussion.[20] Noting that the UN's agendas are already saturated, Chinese participants at the various SIPRI workshops also suggested other forums such as the Brazil–Russia–India–China–South Africa (BRICS) group or the Shanghai Cooperation Organisation (SCO) as viable venues. However, this resulted in questions from US participants as to the ability of such bodies to move beyond political agendas and blocs.

[17] See chapters 13 and 14 in this volume.

[18] See also e.g. Hagström, M., 'Military applications of machine learning and autonomous systems', ed. Boulanin (note 12); and Rickli, J.-M., 'The destabilizing prospects of artificial intelligence for nuclear strategy, deterrence and stability', ed. Boulanin (note 12).

[19] See also e.g. chapters 10 and 11 in this volume.

[20] See also e.g. chapter 11 in this volume.

III. Addressing gaps in AI assumptions and capabilities

Rather than avoiding the issue of differing political agendas and perspectives, the East Asia workshop and this volume have sought to confront differences in perceptions, politics, technologies and militaries head-on. This is intended to facilitate a better understanding of some of the national and regional biases and assumptions that are shaping the impact of AI on strategic stability and nuclear risk at the technical and strategic levels. While each of the countries and viewpoints featured in this volume is unique, there are enough common concerns over the AI stability–instability paradox that there may be greater room to explore how to address the issues of asymmetry, signalling, and intentional and unintentional escalation that are described.

To this end, it is crucial to focus on, and even create a matrix of, what makes each country in East Asia both similar and distinctive in how it addresses the integration of AI into nuclear forces. This could focus on the oft-missed elements of threat perceptions, strategic culture, geography, third-party actors, alliances and non-state actors. Several of the authors in this volume have already offered their own tables and diagrams of how these elements interact at the technical and strategic level.[21] This can be expanded even further by country and region. However, beyond concepts, these variables must be triangulated with the AI-related technologies under development and the platforms being deployed. Doing so would indicate not only how AI algorithms are being built, but also how they fit the nuclear environment and advances within East Asia and beyond.

[21] See also e.g. chapters 10 and 14 in this volume.

About the authors*

Arie Koichi (Japan) is a senior researcher at the National Institute for Defense Studies, Tokyo. He ranks as a lieutenant colonel in the Japan Ground Self-Defense Force (JGSDF) and has served as unit commander and staff throughout Japan. He served in southern Iraq in 2004 as part of the first contingent of the Iraq Reconstruction Support Unit, for which he was embedded as JGSDF liaison officer in the headquarters of Multi-National Division (South East). His area of expertise is nuclear strategy and nuclear deterrence. Since 2012 he has been a part-time lecturer at Graduate School of International Cooperation Studies, Takushoku University, where he teaches US nuclear strategy. He has a master's degree from the National Institution for Academic Degrees and University Education and a doctorate from Takushoku University.

Cai Cuihong (China) is a professor with the Center for American Studies at Fudan University, Shanghai. Prior to this, she worked for the university's Foreign Affairs Office between 1996 and 2001. She was a visiting scholar at the Georgia Institute of Technology, United States, in 2002, and at the University of California, Berkeley, USA, in 2007, as well as an invited fellow in the 2007 programme on US national security sponsored by the US State Department. She is a member of the Shanghai Association of American Studies. She has bachelor's and master's degrees in biophysics and a doctorate in international relations from Fudan University and a bachelor's degree in English language and literature from Shanghai International Studies University.

Hwang Il-Soon (South Korea) is a professor in the Department of Nuclear Engineering at Seoul National University. He is a nuclear energy specialist. As director of the Nuclear Power Performance Research Center, he was responsible for nuclear plant safety and management of ageing, establishing a periodic safety review protocol. While serving as the director of the Nuclear Transmutation Energy Research Center of Korea (NuTREK), he used his experience on the Yucca Mountain Project nuclear waste depository to develop the Proliferation-resistant, Environment-friendly, Accident-tolerant, Continual and Economical Reactor (PEACER) and the PyroGreen system. He chairs or co-chairs the Korean Nuclear Plant Aging Management Network, Nypro Engineering and Technology (NET), Nuclear Power Infrastructure Development, the Informal Heavy Liquid Metal Coolant Interest Group, a joint task force of the Organisation for Economic Co-operation and Development (OECD) and the Nuclear Energy Agency (NEA), an International Atomic Energy Agency (IAEA) consultancy group, the Forum on Climate Change and Energy Policy, and the Summit of Honor on Atoms for Peace and Environment (SHAPE). He has a bachelor's degree from Seoul National University and a doctorate from the Massachusetts Institute of Technology, USA.

* In accordance with the linguistic rules in China, Japan and Korea, the names of authors from these countries appear with the family name first throughout this volume.

Hwang Ji-Hwan (South Korea) is an associate professor at the University of Seoul. His research interests include diplomatic policy and the relationship between North Korea and South Korea. He has a bachelor's degree in diplomacy from Seoul National University, a master's degree in political science from Seoul National University and the University of Colorado, USA, and a doctorate in political science from the University of Colorado.

Jiang Tianjiao (China) is an assistant professor in the Department of International Relations at Shanghai International Studies University. Previously, he was as assistant researcher at the Centre for Cyberspace Governance Studies at Fudan University. He was a visiting scholar at the Sigur Center for Asian Studies, George Washington University, USA, in 2017. He has a doctorate from Fudan University, where he specialized in arms control and regional security.

Vasily Kashin (Russia) is a senior fellow at the Higher School of Economics and the Russian Academy of Science (RAS), Moscow. He is also affiliated with the Russian International Affairs Council and the RAS Institute of Far Eastern Studies. He is an expert on China's military–industrial complex.

Kim Ji-Sun (South Korea) is a professor in the Department of Nuclear Engineering, Seoul National University.

Vadim Kozyulin (Russia) is project director of the Asian Security Project and the Emerging Technologies and Global Security Project at the PIR Center, Moscow. He has served as an expert on political science and as a professor at the Russian Academy of Military Science. He was formerly an officer of the Soviet and then Russian Ministry of Foreign Affairs, worked at the *Moscow News Review* and was later a representative in Russia of Kazspetsexport. In 2000–2002, he completed the programme on management of military and technical cooperation at the Russian Foreign Trade Academy. His research interests include the 2013 Arms Trade Treaty, Russia's military and technical cooperation with foreign countries, and stability in Central Asia and Afghanistan. He is a graduate of Moscow State Institute for International Relations (MGIMO).

Li Xiang (China) is an assistant engineer with the China Shipbuilding Information Center. He has a master's degree from Renmin University, Beijing.

Liu Yangyue (China) is an associate professor in the College of Humanities and Social Sciences of the National University of Defense Technology (NUDT), Changsha. He is also a research fellow in the NUDT College of International Studies and Cultural Security Studies Group. His research interests include Internet politics, Asian politics and international relations. He has a doctorate from the Department of Asian and International Studies of the City University of Hong Kong.

Nishida Michiru (Japan) is head of the Disarmament Unit in the Arms Control and Disarmament Division of the Japanese Ministry of Foreign Affairs (MFA) and a special assistant for disarmament and non-proliferation issues. He previously served as first secretary with the Japanese delegation to the Conference on Disarmament in Geneva, primarily in charge of nuclear and space issues. He was also a visiting associate professor at the Research Center for Nuclear Weapons Abolition of Nagasaki University. Before being posted to Geneva, he served as head of the Export Control Unit of the MFA's Non-Proliferation, Science and Nuclear Energy Division. He was head of the Political Section of the Consulate-General of Japan in Karachi, Pakistan, and a member of the Japanese delegation to the Six-Party Talks in 2005.

Lora Saalman (USA) is an associate senior fellow on armament and disarmament at SIPRI and a senior fellow with the Global Cooperation in Cyberspace Programme of the EastWest Institute, New York. She has also worked as an associate professor at the Daniel K. Inouye Asia–Pacific Center for Security Studies, Honolulu, USA, an associate in the Nuclear Policy Program at the Carnegie–Tsinghua Center for Global Policy, Beijing, an adjunct professor at Tsinghua University, a research associate at the Wisconsin Project on Nuclear Arms Control, and a visiting fellow at the Observer Research Foundation, India, and the James Martin Center for Nonproliferation Studies and an intern at the IAEA, Vienna. She has a bachelor's degree from the University of Chicago and a master's degree with a certificate in non-proliferation from the Monterey Institute of International Studies. She was the first US citizen to earn a doctorate from Tsinghua University's Department of International Relations, completing all her coursework in Chinese.

Su Fei (China) is a researcher within the China and Global Security Programme at SIPRI. Her current research focuses on China's engagement with North Korea, South Korea and Japan. Prior to her current post, she lived and studied in Seoul for three years, where she strengthened her fluency in Korean and obtained a master's degree in public administration with a focus on governance from the Graduate School of Public Administration of Seoul National University, where she wrote her dissertation in Korean.