

Studies in Peace and Security

Series Editors: Christopher Daase · Simone Wisotzki · Anna Leander

Thomas Reinhold
Niklas Schörnig *Editors*

Armament, Arms Control and Artificial Intelligence

The Janus-faced Nature of Machine
Learning in the Military Realm

 Springer

Studies in Peace and Security

Series Editors

Christopher Daase, Peace Research Institute Frankfurt (PRIF), Frankfurt am Main, Germany

Simone Wisotzki, Peace Research Institute Frankfurt (PRIF), Frankfurt am Main, Hessen, Germany

Anna Leander, Department of International Relations and Political Science, Graduate Institute of International and Development Studies, Geneva, Switzerland

● آقای هوش مصنوعی ●

🏢 رسانه هوش مصنوعی دانشگاه تهران 🏢

@MrArtificialintelligence

This book series offers an outlet for innovative research on peace and security in the context of international relations and peace and conflict studies. *Studies in Peace and Security* aims to facilitate the interdisciplinary dialogue on peace and security of diverse academic disciplines. The series features cutting-edge scientific and scholarly studies on causes of international and intra-state conflicts and ways to solve them.

Studies in Peace and Security features monographs, edited volumes as well as text books and handbooks from a variety of disciplines that seek to advance theories and empirical research on peace, conflict and security. Relevant topics include, but are not limited to, international peace and security, arms control and disarmament, international norms, regimes and organizations, international law, military and non-military intervention, peace-building and peace consolidation, democratization as well as radicalization and political violence. Methodologically, the series covers both quantitative as well as qualitative approaches.

Thomas Reinhold • Niklas Schörnig
Editors

Armament, Arms Control and Artificial Intelligence

The Janus-faced Nature of Machine Learning
in the Military Realm

● آقای هوش مصنوعی ●

🏠 رسانه هوش مصنوعی دانشگاه تهران 🏠


@MrArtificialintelligence




 Springer



Editors

Thomas Reinhold 
PEASEC
TU Darmstadt
Darmstadt, Germany

Niklas Schörnig 
Peace Research Institute Frankfurt
Frankfurt am Main, Hessen, Germany

ISSN 2730-9800

ISSN 2730-9819 (electronic)

Studies in Peace and Security

ISBN 978-3-031-11042-9

ISBN 978-3-031-11043-6 (eBook)

<https://doi.org/10.1007/978-3-031-11043-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Acknowledgements

Many books do have a long history, and this book is no exception. The general idea to write a book about “emerging technologies” and arms control developed after PRIF’s annual conference “Verification in Crisis—the Crisis of Verification. New Technology as a Hurdle to and an Enabler of Verification in Arms Control,” organized by Niklas and held late in 2018. It took, however, a little bit over a year until the project really started, now with a clear focus on artificial intelligence and arms control and with a second editor on board: Thomas. Joining the project in February 2020, the project came to a grinding halt only a few weeks later due to the Covid-19 pandemic.

As anyone who has worked on a major project with colleagues from different countries during the pandemic knows, conditions were even more difficult—from homeschooling to the fact that some colleagues first had to set up an office at home or simply fell ill. Nevertheless, we are very happy that no one had to leave due to illness. We are all the more pleased to have finished this book.

We would like to first thank all our wonderful colleagues who not only provided their valuable time and knowledge but who often went the extra mile under difficult conditions when we once again asked to clarify a point here or add an argument there. Without their expertise and patience, the book would have never seen the light of day.

The same heartfelt gratitude goes to our families for their constant support during the writing and editing of this book with the Covid conditions putting more strain on our families than usual.

Yet, finishing the book would not have been possible with the help and support of many other people. We want to thank Anna Leander, Simone Wisotzki, and Christopher Daase for giving us the opportunity to be the first book published in their new “Studies in Peace and Security”-series. This project was supported by Professor Christian Reuter and his team and by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

Friedjof Meyer assisted us in our own research, while Selma Mustafić, Frank Kuhn, Anja-Liisa Gonsior, and Hannah Appich supported us getting the manuscripts in the right shape.

Matthew Harris was a great help to us as our language editor. He managed to walk a fine line between keeping everyone's individual style and being thorough when it came to mistakes, and it goes without saying that any remaining error is, of course, our fault.

Finally, we have to thank our external reviewer for his critical yet always constructive comments. Thanks to his eagle eyes, every text gained enormously in quality.

Contents

Introduction	1
Niklas Schörnig and Thomas Reinhold	
Introduction into Artificial Intelligence and Machine Learning	11
Melanie Reuter-Oppermann and Peter Buxmann	
The Military Rationale for AI	27
Frank Sauer	
Military AI Applications: A Cross-Country Comparison of Emerging Capabilities	39
Sophie-Charlotte Fischer	
Artificial Intelligence as an Arms Control Tool: Opportunities and Challenges	57
Niklas Schörnig	
Verifying the Prohibition of Chemical Weapons in a Digitalized World	73
Alexander Kelle and Jonathan E. Forman	
AI and Biological Weapons	91
Filippa Lentzos	
Doomsday Machines? Nukes, Nuclear Verification and Artificial Intelligence	101
Jana Baldus	
AI, WMD and Arms Control: The Case of Nuclear Testing	117
Anna Heise	
Artificial Intelligence in Conventional Arms Control and Military Confidence-Building	129
Benjamin Schaller	

Cyber Weapons and Artificial Intelligence: Impact, Influence and the Challenges for Arms Control 145
Thomas Reinhold and Christian Reuter

Drones and Lethal Autonomous Weapon Systems 159
Anja Dahlmann

No, Not That Verification: Challenges Posed by Testing, Evaluation, Validation and Verification of Artificial Intelligence in Weapon Systems 175
Maaïke Verbruggen

Applying Export Controls to AI: Current Coverage and Potential Future Controls 193
Kolja Brockmann

Arms Control for Artificial Intelligence 211
Thomas Reinhold

Editors and Contributors

About the Editors



Thomas Reinhold is a research associate and Ph.D. student at the Department of Science and Technology for Peace and Security (PEASEC) at TU Darmstadt. He is working on IT-based measures for arms control of military activities in cyberspace as well as the challenges of the militarization of artificial intelligence.



Niklas Schörnig is a senior researcher at the Peace Research Institute Frankfurt (PRIF) and a head of PRIF's research group on Emerging Technologies, Order and Stability (rETOS). His main research focusses on the military use of robots and AI, arms dynamics and arms control, and military ethics. He is the editor of the EU Non-Proliferation and Disarmament eLearning course (<https://nonproliferation-elearning.eu/>) and can be followed on Twitter (@NiklasSchoernig).

Contributors



Anna Heise has a Ph.D. in physics from the University of Hamburg. She wrote her thesis at the Carl Friedrich von Weizsäcker Centre for Science and Peace Research about nuclear test verification and is a certified Data Analyst. Currently, she teaches both at school and at university.



Jana Baldus is a research associate in the field of international security and nuclear weapons at the Peace Research Institute Frankfurt. In addition to her interest in the implications of the use of new and emerging technologies in the nuclear field, she is particularly interested in the role of divergent perceptions about the nuclear order and how these contribute to resistance and contestation to the nuclear nonproliferation regime.



Kolja Brockmann is a Senior Researcher in the Dual-use and Arms Trade Control Programme at the Stockholm International Peace Research Institute (SIPRI). He holds a master's degree with distinction in Non-Proliferation and International Security from King's College London. Prior to joining SIPRI in 2017, he interned with the German Federal Office for Economic Affairs and Export Control (BAFA). He conducts research in the fields of export control, nonproliferation, technology governance and sanctions, with a particular focus on the multilateral export control regimes.



Peter Buxmann is a Full Professor and head of the Software & Digital Business Group at the Technical University of Darmstadt. He is also a member of numerous supervisory and management bodies, including the advisory board of the Weizenbaum Institute for the Networked Society—the Internet Institute in Berlin, and the supervisory board of Eckelmann AG, Wiesbaden, where he is responsible for digital transformation.

One focus of his work is the transfer to business. He has given speeches at many events, including the Software Business Conference at MIT, ECM World in Düsseldorf, the Hamburg IT Days and the Frankfurt IT Days. Furthermore, he is dedicated to business education, particularly in the areas of digital transformation and artificial intelligence. Peter Buxmann has co-founded two companies, supports a large number of start-ups, and is a Chairman of the advisory board of the HIGHEST innovation and start-up center at the Technical University of Darmstadt—after managing and helping to establish the center for five years. He is also a member of the steering committee of the TechQuartier in Frankfurt. Together with Holger Schmidt, he hosts the FAZ podcast “Artificial Intelligence.”

His research focuses on the digitalization of businesses and society, methods and applications of artificial intelligence, and the tension between data economics and privacy. He is the author of more than 300 publications, which have appeared in *Frankfurter Allgemeine Zeitung*, *Süddeutsche Zeitung*, and in international journals (e.g., *Management Information Systems Quarterly*, *Information Systems Research*, *Journal of Information Technology*, *European Journal on Information Systems*, *Information Systems Journal*) and conference proceedings (e.g., *International Conference on Information Systems* and *European Conference on Information Systems*).

Peter Buxmann was born in Frankfurt in 1964. He studied economics with a focus on information systems at the University of Frankfurt, where he also received his doctorate. He habilitated after a research and teaching period at the Haas School of Business at the University of California in Berkeley. From 2000 to 2004, he was a Professor of Information Systems and Information Economics at Technische Universität Freiburg, before transferring to the Technical University of Darmstadt.



Anja Dahlmann is the head of the Berlin Office of the Institute for Peace Research and Security Policy at the University of Hamburg (IFSH) and a researcher, focusing on arms control and autonomous weapon systems. Previously, Anja Dahlmann worked at the German Institute for International and Security Affairs and was the head of the International Panel on the Regulation of Autonomous Weapons (iPRAW). She holds a Master's Degree in Political Science from the University of Göttingen.



Sophie-Charlotte Fischer is a senior researcher at the Center for Security Studies at ETH Zurich. Her research focuses on the geopolitics and governance of emerging technologies with a focus on AI. Previously, she worked with the German Foreign Office, the UN Office for Disarmament Affairs, and NATO Defense College. She holds a doctorate in Political Science from ETH Zurich, an MA in International Security from Sciences Po Paris, and a BA in Liberal Arts from University College Maastricht.

Alexander Kelle is a Senior Researcher at the Berlin Office of the Institute for Peace Research and Security Policy (IFSH). He was a Senior Policy Officer in the Office of Strategy and Policy at the Organisation for the Prohibition of Chemical Weapons (OPCW) from 2013 to 2019.

Jonathan E. Forman was the OPCW Science Policy Advisor from 2013 to 2020. The views expressed in his chapter are the authors' own and should not be attributed to any past or current employer.



Filippa Lentzos is a Senior Lecturer in Science & International Security at the Department of War Studies and a Co-Director of the Centre for Science and Security Studies (CSSS) at King's College London. She is also an Associate Senior Researcher at the Stockholm International Peace Research Institute (SIPRI) and a Non-Resident Scholar at the James Martin Center for Nonproliferation Studies (CNS).



Christian Reuter is a Professor at the Technical University of Darmstadt. His chair Science and Technology for Peace and Security (PEASEC) in the Department of Computer Science combines computer science with peace and security research. On the intersection of the disciplines (A) Cyber Security and Privacy, (B) Peace and Conflict Studies as well as (C) Human–Computer Interaction, he and his team specifically address: (1) Peace Informatics and technical Peace Research, (2) Crisis Informatics and Information Warfare as well as (3) Usable Safety, Security and Privacy.



Melanie Reuter-Oppermann is a postdoctoral researcher in the Software & Digital Business Group at the Technical University of Darmstadt. In 2017, she received her Ph.D. in Operations Research from the Karlsruhe Institute of Technology (KIT) on the analysis and optimization of Emergency Medical Services (EMS) systems. At KIT, she established the HealthCareLab at the Karlsruhe Service Research Institute. She is a joint coordinator of the European Working Group on Operational Research Applied to Health Services (ORAHS). In her research, she applies Information Systems and Operations Research methods to support decision-making in healthcare. Besides EMS, her recent research interest is in primary care services, hospital and blood logistics as well as crisis management. She recently received the Julius von Haast Fellowship from the Royal Society of New Zealand.



Frank Sauer is the Head of Research at the Metis Institute for Strategy and Foresight as well as a Senior Research Fellow at the Bundeswehr University Munich. The main foci of his research are nuclear issues and emerging technologies. Frank also serves on various expert panels in an advisory role for governments, civil society, and industry. He co-hosts the German language podcast “Sicherheitshalber” on all things security and defense. You can follow Frank on Twitter @drfranksauer.



Benjamin Schaller is a former research fellow and Ph.D. graduate from UiT—The Arctic University of Norway. His research focused on trust and distrust in defense and security politics, NATO-Russia relations, Arctic security, arms control, and military confidence-building. In connection with the German OSCE chairmanship 2016, he worked as a desk officer for arms control and military confidence-building at the German Federal Foreign Office. Between 2020 and 2021, he was part of the first cohort of the Harvard Arms Control Negotiation Academy (ACONA).



Maaïke Verbruggen Originally a historian and a sociologist, Maaïke Verbruggen now studies the future of warfare. Her specialty is the interplay between emerging technologies, military innovation, and arms control. She is a Doctoral Researcher at the Center for Security, Diplomacy and Strategy (CSDS) at the Brussels School of Governance (BSOG) at the Vrije Universiteit Brussel, Belgium. She is currently finishing up her Ph.D. on military innovation in Artificial Intelligence—and particularly the controversies behind its long history.

Abbreviations

ACFE	Adapted Treaty on Conventional Armed Forces in Europe
AI	Artificial intelligence
AEMI	Annual Exchange of Military Information
AGI	Artificial General Intelligence
ANN	Artificial neural networks
API	Application Programming Interface
ASCI	Accelerated Strategic Computing Initiative
ATLAS	Advanced Targeting and Lethality Automated System
AWS	Autonomous weapon systems
BWC	Biological Weapons Convention
BOLT	Broad Operational Language Translation
CBM	Confidence-Building Measures
CBN	Chemical, biological, nuclear
CCIAD	Coordination Cell of Defense Artificial Intelligence
CCW	United Nations Convention on Certain Conventional Weapons
CFE	The Treaty on Conventional Armed Forces in Europe
CI	Challenge Inspection
CM	Confusion matrix
CPU	Central Processing Unit
CSBM	Confidence- and Security-Building Measures
CSCE	Conference on Security and Co-operation in Europe
CSP	Conference of State Parties
CTBT	Comprehensive Nuclear-Test-Ban Treaty
CTBTO	Comprehensive Nuclear-Test-Ban Treaty Organisation
CW	Chemical Weapons
CWC	Chemical Weapons Convention
DARHT	Dual-Axis Radiographic Hydrotest Facility
DARPA	Defense Advanced Research Projects Agency
DAT	Declaration Assessment Team
DIA	Defense Innovation Agency

DIU	Defense Innovation Unit
DL	Deep Learning
DNA	Deoxyribonucleic acid
DoD	Department of Defense
EDA	European Defense Agency
EDIS	Electronic Declaration System
EU	European Union
FCAS	Future Combat Air System
FFM	Fact-Finding Mission
GARD	Guaranteeing AI Robustness Against Deception
GOFAI	“good old-fashioned” Artificial Intelligence
HSI	Hyperspectral and Full Spectrum Imaging
IAEA	International Atomic Energy Agency
IAU	Investigation of alleged use of CW
IBM	International Business Machines Corporation
ICRC	International Committee of the Red Cross
IDC	International Data Center
IDF	Israel Defense Forces
IEDs	Improvised Explosive Devices
IHL	International humanitarian law
IHRL	International Human Rights LAW
IMS	International Monitoring System
INF	Intermediate-Range Nuclear Forces
INS	Inertial Navigation Systems
IPR	Intellectual Property Rights
iPRAW	International Panel on the Regulation of Autonomous Weapons
IoT	Internet of Things
ISR	Intelligence, Surveillance and Reconnaissance
IT	Information Technology
ITT	Intangible Transfers of Technology
JADC2	Joint-All-Domain Command-and-Control-Concept
JAIC	Joint Artificial Intelligence Center
JIM	Joint Investigative Mechanism
LAWS	Lethal Autonomous Weapon System
LIDAR	Light Detection and Ranging
LISP	Abbreviation for “List Processing”
M&S	Modeling and Simulation
MADCAT	Multilingual Automatic Document Classification, Analysis and Translation
MBFR	Mutual and Balanced Force Reductions
ML	Machine Learning
MTCR	The Missile Technology Control Regime
NATO	North Atlantic Treaty Organisation
NC3	Nuclear command, control and communication

NGO	Non-Governmental Organization
NIF	National Ignition Facility
NPT	Treaty on the Non-Proliferation of Nuclear Weapons
NSA	National Security Agency
NSG	Nuclear Suppliers Group
NTM	National Technical Means
OECD	Organization for Economic Co-operation and Development
OPCW	Organization for the Prohibition of Chemical Weapons
OS	Open Skies
OSCE	Organization for Security and Cooperation in Europe
PEASEC	Science and Technology for Peace and Security
PLA	Chinese People's Liberation Army
PTBT	Partial Test Ban Treaty
RADAR	Radio Detecting and Ranging
RNA	Ribonucleic acid
REB	Reviewed Event Bulletin
R&D	Research and Development
S&T	Science and Technology
SAB	Scientific Advisory Board
SALT	Strategic Arms Limitation Talks
SAR	Synthetic Aperture Radar
SEL	Standard Event List
SIFT	Scale-invariant Feature Transform
SIPRI	Stockholm International Peace Research Institute
SIGINT	Signals Intelligence
SIX	Secure Information Exchange
SSP	Stockpile Stewardship Program
TEV&V	Testing, Evaluation, Validation and Verification
TWG	Temporary Working Group
UAV	Unmanned Aerial Vehicles
UCAV	Unmanned Combat Aerial Vehicles
UAV/UUV	Weaponized uncrewed aerial or underwater vehicles
UN	United Nations
UN(O)	United Nations (Organization)
UK	United Kingdom
US	United States
US(A)	United States (of America)
vDEC	virtual Data Exploitation Centre
Vdoc	Vienna Document on Confidence- and Security-Building Measures
V&V	Validation and Verification
WA	Wassenaar Arrangement (on Export Controls for Conventional Arms and Dual-Use Goods and Technologies)
WMD	Weapon(s) of Mass Destruction
XAI	Explainable Artificial Intelligence

Introduction



Niklas Schörnig and Thomas Reinhold

● آقای هوش مصنوعی ●

🏢 رسانه هوش مصنوعی دانشگاه تهران 🏢

@MrArtificialintelligence

Abstract In 1987, Allan Din published the seminal book “Arms and Artificial Intelligence,” in which he argued that the future military use of AI would be a double-edged sword. Warning about control failures and accidental war on one hand, Din also pointed out the potential of AI to enhance arms control. 35 years later, what was a niche technology in Din’s day has since become one of the most influential technologies in both the civilian and military sectors. In addition, AI has evolved from sophisticated yet deterministic expert systems to machine learning algorithms. Today, AI is about to be introduced in almost every branch of the military, with a variety of implications for arms control. This book reflects the work of the individual authors and identifies common themes and areas where AI can be used for the greater good or where its use calls for particular vigilance. It offers an essential primer for interested readers, while also encouraging experts from the arms control community to dig more deeply into the issues.

1 The Use of AI as a Revolution in Military Affairs

“The envisaged uses of computer and IT techniques in weapon systems give rise to both skepticism and concern, for example because of the risk of control failures leading to crisis and accidental war. There are, however, also possible applications of these techniques within arms control which may have a more positive connotation” (Din, 1987, p. 8). When Allan Din wrote these words in his seminal edited volume “Arms and Artificial Intelligence” in 1987 they almost sounded like science fiction. However, when Din and his co-authors talked about AI they had an understanding of it that is completely different how it is perceived today. In the 1980s, AI often boiled

N. Schörnig (✉)
Peace Research Institute Frankfurt, Germany
e-mail: schoernig@hsfk.de

T. Reinhold
PEASEC, TU Darmstadt, Darmstadt, Germany
e-mail: reinhold@peasec.tu-darmstadt.de

down to so-called expert systems, that is, highly complex, deterministic systems supporting decision-making “based on the heuristic knowledge of domain experts in combination with decision rules” (Orhaug, 1987, p. 167), or problem solving by brute force, a rather simple trial-and-error approach limited by processing power.

Today, things look rather different: Not only has the processing power of CPUs increased by a factor of 1000¹—following Moore’s Law until very recently—new AI techniques such as machine learning have revolutionized the potential of AI applications. When Stanley, the autonomous Volkswagen Touareg, won the DARPA Grand Challenge in 2005 it was the first of five cars to cover the 213 km distance without an accident—in the open desert, without any other traffic. Even 17 years later, fully self-driving cars are still not on the market, but the assistance systems are still aiming at making the driver in the cockpit almost superfluous and already work quite well in well manageable situations.

Some experts claim, and rightly so, that development will continue at a rapid pace, especially as AI is, as Paul Scharre puts it, “not a discrete technology like a fighter jet or locomotive, but rather is a general-purpose enabling technology, like electricity, computers, or the internal combustion engine” (Scharre, 2019). Since civilian advances in AI have a high dual-use character, they have also advanced the military use of AI to a massive degree. Other experts are less optimistic and warn that AI is still best applied to very specific tasks and that the vision of a more universal, flexible and adaptable AI might turn out to be a dead end.

But so far AI seems to provide the technology for enhancing solutions for technical challenges and the latest technical advancements in computer processing power and size described above now allow powerful devices that can handle the processing of AI algorithms, making it increasingly clear that AI will permeate and transform all military domains, from reconnaissance and analysis to key decision-making processes on both the tactical as well as the strategic level, and, finally, the direct execution of a military strike or attack. There are numerous examples, mostly from the US military, that is (still) at the forefront of implementing AI in the military: The US Navy, for example, is “looking to leverage advanced technological capabilities in artificial intelligence (AI) and machine learning (ML) in the tactical and operational realm” (Munoz, 2022, p. 8). In the future, a new Joint-All-Domain Command-and-Control-Concept (JADC2) is expected to combine data from a multitude of sensors and apply AI-enhanced evaluation to the data in order to identify targets and recommend the optimal weapon (White, 2021, p. 21).

While military AI still consists of isolated islands in many places, no-one would doubt that it will have an even stronger impact in the years to come, when the so-called “Internet of Military Things” will “change the landscape of defense operations,” as the trade magazine *Jane’s Defence Weekly* predicts (Torruella, 2021, p. 3).

¹In 1985, Intel’s new 80,386 CPU combined 275,000 transistors on one chip, while the current generation of microprocessors squeeze over 3 billion transistors into a very small space. However, the number of transistors is of course not the only factor determining a CPU’s performance.

In any case the steady rise in the use of military AI is leading to an ever-increasing acceleration of decision processes and cycles. It is no wonder that many experts, some of whom are also represented in this volume, have argued for some time that the introduction of AI will primarily lead to an acceleration of military actions and responses, shorter reaction times and higher alert levels. This assessment is also shared by the military. The German armed forces, for example, expect a “battle at machine speed” in the future, with decisions to be made in minutes or even seconds, rather than hours (Doll & Schiller, 2019, p. 4).

Thus, while military commanders recognize a tremendous advantage when they—and only they—have a significant advantage in the use of AI, arms controllers and other critics primarily see the dangers of an unhindered and unrestricted military “AI race.”

2 The Purpose of the Book

However, such observations are neither new nor innovative, and debating only the military impact of AI on war in general would probably not warrant another book. Our book seeks to go a step further and look at the military use of AI from the perspective of arms control and verification. While we describe the idea behind arms control and verification in more detail later in the book, it is fair to say that all serious arms control needs verification to ensure compliance and be effective. Unfortunately, when it comes to arms control, AI causes potential problems unseen in the older days of physical weapon systems. Not only is arms control hindered by the fact that in contrast to hardware such as tanks, planes, or missiles, which can be physically inspected and counted, software code is notoriously hard to control and verify—if ever. Given the large capacity of modern memory hardware, even extremely complex programs can be stored on fingernail-size memory cards. Software can be updated or replaced in an instant—even if a specific military system passes an inspection, the chance is high that a software update will increase its performance and the dangers it poses tremendously. Consequently, AI will have a very sizable impact on arms control—for better or worse. Unfortunately, many arms control experts who are very familiar with the particular weapon category they work on, still shy away from dealing with AI as they fear that their knowledge in computer science is not sufficient. Our book thus aims at broadening understanding of the relevance of software, AI and ML in the military and arms control realm and seeks to encourage experts to look more deeply into the advantages and disadvantages of AI in their field. The book offers background knowledge about what AI and ML are, how they work, and what they can and cannot achieve, and provides both broader perspectives on the way AI will transform the military as well as insights into key players. It also offers an overview of the relevance of software, AI and ML within several weapon fields in the realm of nuclear, biological and chemical (NBC) weapons, conventional weapons and emerging technologies and examines how the respective fields are dealing with the increasing relevance of new AI-technologies.

For those who are more strongly focused on AI, the book introduces relevant theoretical concepts of arms control and verification, and the way different AI developments will impact arms control. While almost all chapters could easily be twice as long as they are now, all authors were asked to be brief, crisp, and understandable. Consequently, the chapters are not only usable for gaining knowledge but are also very suitable for classroom teaching.

3 The Structure of the Book

After this introduction, the book is divided into three sections. The first section contains theoretical reflections and looks at key actors in the field. The section starts with a text by Peter Buxmann and Melanie Reuter-Oppermann. They provide an informed introduction to the topic of artificial intelligence and machine learning. Without drifting into formal or mathematical argumentation (which has been placed in the appendix), they first provide a short historical overview of artificial intelligence and machine learning followed by a more concrete introduction to different forms of machine learning algorithms and methods for measuring algorithm quality. This chapter is unrelated to the topics of armament and arms control and can be used as a general introduction to AI. Chapter three, written by Frank Sauer, describes the military rationale for the use of AI. Sauer starts with an understanding of AI encompassing automated tasks “which previously required the application of human intelligence” (p. 27). While this understanding is rather broad, it is also common within military circles. Sauer concludes that in the debate on the military use of AI and ML both are “simultaneously over- and underestimated” (p. 27), blurring clear-cut discourse on opportunities and threats. Frank also sees signs of a dynamic that is detached from actual military needs, arguing that many military officials are employing AI in the armed forces “because everyone else is doing it” (p. 28) and pointing to the pressure many militaries see themselves under. But he also concludes that AI has much to offer the military, at least at first sight and from a strictly military point of view where AI allows faster targeting cycles which eventually lead to superiority over the opponent. He concludes that “the hype is real” but cautions that “so are the risks” (p. 28), and argues that the widespread misunderstanding of AI’s strengths and weaknesses is in large part responsible for making the introduction of AI in military applications fraught with risk due the hype surrounding it (p. 35).

Chapter four, written by Sophie-Charlotte Fischer, looks more closely at issues from the end of Sauer’s more general chapter and introduces the key players regarded as responsible for the AI arms race. Fischer argues “that important clues to inform the nascent academic and policy discourse on the military and broader security effects of AI can be derived from analyzing and comparing how different countries pursue military AI—in what kind of applications they invest and in which selected areas they already deploy AI” (p. 40). After developing a framework for analysis, Fischer assesses the military capabilities of four countries, the United

States, China, France, and Israel. She concludes that all four countries view synergies between the commercial and military sector as critical to realizing their AI objectives and are in the process of implementing AI “across a wide range of areas including logistics and training, cyber and information operations, Intelligence, Surveillance and Reconnaissance (ISR), and (semi-)autonomous vehicles as well as command and control” (p. 52). However, each country differ significantly in their approach to the risks associated with the application of AI in the military.

In the fifth chapter, Niklas Schörnig looks at the brighter side of the use of AI in a military context and examines how AI can be used to foster arms control. After a general overview presenting the theoretical background of arms control, disarmament and non-proliferation from the specific perspective of verification, Schörnig systematizes the use of AI for arms control in several broader categories, including the use of AI for translation and analysis of text in arms control and verification contexts, the analysis of graphical data, other sensory data, and multimodal data. He concludes that while AI will not replace inspectors in the foreseeable future, it nevertheless offers very helpful support that facilitates the work of inspectors and should be used more in the future.

The second major section of the book, “Empirical Examples from Different Fields of Arms Control,” starts with chapter six written by Alex Kelle and Jonathan E. Forman, both of whom have a background as former employees of the Organization for the Prohibition of Chemical Weapons (OPCW). They address the issue of “Verifying the Prohibition of Chemical Weapons in a Digitalized World.” In order to understand how new technologies including AI fit into the elaborate verification mechanism of the OPCW, the text first offers a basic understanding of the different verification rules and procedures implemented by the OPCW. They also show that the use of state-of-the-art science and technology for verification purposes flows directly from the Chemical Weapons Convention itself. While AI can enhance verification, the authors also draw attention to the profound changes through which the chemical industry has gone in recent years due the adoption of AI as part of the so-called Industry 4.0. They conclude that this is no time to be afraid of the rapid changes in science and technology, but that scientific literacy is the key to keeping track of both beneficial and malicious use.

In chapter seven Filippa Lentzos looks at AI and biological weapons and highlights key impacts of machine learning and automation on biological research, medicine and healthcare. Lentzos argues that these developments could make the production of biological weapons easier and proliferation more likely. She continues that even though biological weapons are completely prohibited by the Biological Weapons Convention, artificial intelligence and other converging technologies are radically transforming the dual-use nature of biology and present significant challenges for the treaty. She discusses these challenges and presents a vision of how biological arms control can evolve in order to remain relevant in the Fourth Industrial Revolution.

Chapter eight, written by Jana Baldus, is the first of two chapters to look at AI and nuclear weapons. Baldus looks, first, at the connection between AI, nuclear weapons and autonomy and points out that during the Cold War earlier forms of AI were quite

common in the nuclear domain. She argues that the use of AI and ML could lead to more reliable early warning and nuclear command systems, generally enhancing nuclear stability. She also points to the downsides, however, including, among others, biased datasets or even increased skepticism toward a high degree of technologization due to the excessive destructiveness of nuclear weapons. She also points out how “AI could help improve the cross-analysis of ISR data, for example to help control treaty declarations” or support the efforts against nuclear proliferation. Like others in this book, Baldus argues that experts in the weapon systems under consideration need to gain an even better understanding of what AI already exists and where and keep track of how these developments will impact nuclear strategy.

The next text, chapter nine, by Anna Heise, delves into an aspect Jana Baldus only touched on: The use of AI in nuclear testing, that is the simulation of nuclear explosions on powerful computer systems. Based on the little that is publicly known about the subject, Heise describes how AI has improved virtual testing and thus avoiding “live” tests with actual nuclear weapons. Heise stresses the human factor and argues that the results of tests “are only as good as the data and models you give them and the knowledge and experience of the person doing the calculations” (???). On this basis she concludes that the future use of AI in testing will “not only be dependent on the technology but on the emotional attitude of those in charge” (???). Heise then looks into the processes of detecting nuclear tests as it has been carried out by the Comprehensive Nuclear-Test-Ban Treaty Organisation (CTBTO) since its foundation in 1996. She explains, for example, how AI can be used to detect tests with seismic wave-form analysis or how AI can help estimate yields of nuclear explosions. Finally, Heise looks at the dangers, emphasized by some observers, such as the analysis of explosions potentially generating proliferation-relevant information on, for example, the design of warheads. She finally concludes that there is already relevant technology for both virtual testing and detection of real nuclear tests, but that these technologies are only being implemented tentatively. Obviously, there is still a lack of trust when it comes to the use of AI in such security-relevant contexts.

With chapter ten, written by Benjamin Schaller, the focus shifts from weapons of mass destruction to conventional aspects of arms control. Based on well-known theories of international relations, Schaller presents the need for conventional arms control and starts with a short overview of European conventional arms control. The European focus may surprise the casual reader, but in fact Europe is the only region in the world where, at least until recently, there was a comprehensive and established arms control architecture in place. Schaller first discusses whether the balance of power will be altered by the use of military AI. He argues that AI will make it even harder to come up with a “balance of power” as quantitative factors become less relevant in contrast to qualitative factors, which are harder to establish. Schaller also argues that at least within the OSCE, the Organization for Security and Cooperation in Europe, AI has played only a minor role, arguing that current differences have caused too many problems for the implementation of AI in fostering arms control to be considered. But he sees chances of fostering conventional arms control, for example by analyzing military information that has been exchanged in the context

of confidence and security building measures, but also by enhancing more concrete verification measures. Schaller concludes by emphasizing what other authors have previously stressed: the importance of maintaining the “human factor” in arms control.

Leaving physical weapons altogether, chapter eleven, written by Thomas Reinhold and Christian Reuter, focuses on “cyber weapons and AI.” After an insightful introduction to the militarization of cyberspace, Reuter and Reinhold examine how the development of future cyber weapons will be influenced and driven by AI and ML. The authors argue that cyber and AI/ML are closely related and that all positive effects of AI and ML on developing software when transferred to the cyber sphere as well as current software architecture of course provides an ideal platform for having AI/ML components added to them. They argue that the problems normally associated with AI, namely the loss of human control due to ever-shorter reaction times, are particularly relevant in the cyber domain, “an environment that is marked by extremely low response times.” Reuter and Reinhold also draw attention to the fact that the black-box character of AI and ML systems could lead to new problems regarding attribution of attacks. But they also see a bright side, for example a time when AI-enhanced algorithms will be able to find slightly altered code instead of looking for exact matches or reveal hackers by identifying their particular “digital fingerprint.”

Many of the previous texts have described lethal autonomous weapons as a prime example of the future use of military AI. In chapter twelve, Anja Dahlmann finally looks at the two most prominent “emerging military technologies,” drones and lethal autonomous weapon systems (LAWS). Dahlmann describes remotely piloted military drones as a step toward autonomy. From a military perspective, future drone systems will probably involve more new functions be carried out autonomously, such as air-to-air combat or manned-unmanned teaming. More autonomy will also offset current shortcomings, such as latency problems or broken or jammed communication links. Dahlmann raises the point that all these autonomous functions will most probably be based on AI and ML, drawing a direct line between current drones and future LAWS. Dahlmann continues to argue that this development will necessitate a new perspective on arms control, with a focus on the element of human control. In that context, Dahlmann also reminds us that many of the components of LAWS will be dual-use. She concludes that, due to the lack of concrete regulation of LAWS, it is only hypothetical whether AI could have a positive impact on arms control for LAWS—and whether only LAWS should be equipped with “some sort of ethical behavior” (???)

The third and last section of this book focuses on the question of “what should be done.” In chapter thirteen, Maaïke Verbruggen focuses on the technical aspects of making ML-based AI reliable. Using the term “verification” in the strict technical sense of software engineering rather than in the sense of arms control, Verbruggen shows the great difficulties when applying time-proven concepts of engineering to software in general and self-learning software in particular. These problems are compounded by the fact that AI is often integrated modularly, leaving open questions of how the AI and the rest of the software interact. She proposes that integration

of verification and validation measures should be structurally integrated into the design process of AI-based software from the very start, arguing for a “correct by construction” approach. While on the one hand Verbruggen stresses that these problems are already being examined by defense ministries around the world, she also fears that establishing international validation and verification standards will become a very difficult task.

In chapter fourteen, Kolja Brockmann discusses how current export control regimes are already applicable to AI and ML algorithms and how they should be improved to restrict the proliferation of malicious AI applications. Brockmann starts from the assumption that there is “lack of clarity about the extent to which export control instruments already cover dual-use goods and technologies used in AI and its military applications” (???). While examining existing export control regimes for dual-use goods, such as the Wassenaar Arrangement, in detail, Brockmann identifies both controls relevant to hardware (e.g., CPUs specifically designed for AI) as well as software, or even “technologies,” understood as specific information necessary for the development of AI tools. He then describes current review processes by, for example, the United States or the European Union and how these processes deal with emerging technologies. Going beyond existing regimes, Brockmann finally looks at challenges and opportunities in applying export controls to AI, weighing up the conflicting aims of export control and describing opportunities and benefits. He concludes that coordination and exchange between the major stakeholders will be the key to finding the right balance in the control of AI exports.

In the final chapter, chapter fifteen, Thomas Reinhold looks at a topic most people would consider a non-starter: the application of hard arms control measures to artificial intelligence and machine learning. While many observers would argue that conventional arms control instruments, such as verification and inspections, cannot be applied to software at all and that only weaker normative restrictions have a chance of being applied, Reinhold looks at best practices from the cyber realm as a source of innovative ideas. To achieve this he disaggregates the process of building an AI application into several independent elements, including training data, classifiers, the AI model and the effectors where the AI is finally applied, and discusses how specifically tailored arms control instruments could be applied separately. Reinhold himself points out that these considerations are currently only theoretical and that significantly more work is required in order to arrive at initial proofs of concept. Viewed optimistically, however, the chapter shows that the statement that hard arms control cannot be transferred to “soft” software needs to be reconsidered.

4 Conclusion

Looking at all fifteen chapters, several general conclusions can be drawn. As was to be expected, AI has an impact on almost all types of weapons. Even if individual weapons are not always optimized by AI, “mosaic warfare” (Torruella, 2021), that is, the enormous relevance of data and information exchange and analysis, has already

arrived in many areas of the military. In more and more instances, humans are supported and assisted by AI, leaving the human as the slowest link in military decision-making. Developments are often driven by the AI race in the civilian sector. The states with a dynamic civilian technological AI base are also the states that want to reap the benefits for the military. Almost all authors fear that the use of military AI will lead to an increased speed of military operations and the need to act faster in times of crisis, leading to instability and hair-trigger alerts. The general unpredictability of current black-box AI algorithms must also be added to this, potentially worsening situations where human soldiers have to trust their computer. Thus, both finding ways to increase the reliability of AI as well as forms of control for AI are the imperatives of future research. But there are also positive developments: In many contexts, projects are exploring how AI can be used to enhance arms control in general and verification in particular. International institutions such as the IAEA are looking very closely at how they can harness AI for their own purposes (IAEA, 2020). While arms control is in its most severe crisis since its introduction in the 1960s, reliable AI might be a key to restarting arms control in a new and reliable fashion. However, there is also agreement that verification should not be outsourced to computers completely, but that AI should primarily aim at supporting human inspectors rather than replacing them.

Finally, we hope that this book will encourage experts from the arms control community who until now have shied away from the topic of artificial intelligence in their respective fields to dig more deeply into the issues. What is needed is genuine interdisciplinarity, something which is far too rarely seen. We hope that our book shows that interaction between the two professions is needed and possible.

References

- Din, A. M. (Ed.). (1987). *Arms and artificial intelligence*. Oxford University Press.
- Doll, T., & Schiller, T. (2019). *Artificial intelligence in land forces* (Position Paper). German army concepts and capabilities development Centre. Retrieved from Cologne: <https://www.bundeswehr.de/resource/blob/156026/3f03afe6a20c35d07b0ff56aa8d04878/download-positionspapier-englische-version-data.pdf>
- IAEA. (2020). *Emerging technologies workshop. Insights and actionable ideas for key safeguard challenges*. (Workshop Report STR-397). IAEA Safeguards.
- Munoz, C. (2022). USN developing AI to drive tactical, operational focus. *Janes Defence Weekly*, 59(3), 8.
- Orhaug, T. (1987). Computer applications in monitoring and verification technologies. In A. M. Din (Ed.), *Arms and artificial intelligence* (pp. 165–178). Oxford University Press.
- Scharre, P. (2019). *Military applications of artificial intelligence: Potential risks to international peace and security*. Center for a New American Security. Retrieved from <https://stanleycenter.org/wp-content/uploads/2020/05/MilitaryApplicationsofArtificialIntelligence-US.pdf>. from Stanley Center <https://stanleycenter.org/wp-content/uploads/2020/05/MilitaryApplicationsofArtificialIntelligence-US.pdf>
- Torruella, A. (2021). Mosaic warfare. *Janes Defence Weekly*, 58(40), 20–25.
- White, A. (2021). Evolving connections. *Janes Defence Weekly*, 58(46), 20–25.

Introduction into Artificial Intelligence and Machine Learning



Melanie Reuter-Oppermann  and Peter Buxmann 

● آقای هوش مصنوعی ●

🏢 رسانه هوش مصنوعی دانشگاه تهران 🏢

@MrArtificialintelligence

Abstract Artificial intelligence (AI) has become an important topic in research as well as industry since its birth in the 1950s. Research on early approaches of machine learning has actually been going on since the 1940s. Still, the question often arises what is actually meant by AI, especially in practice. In this chapter, we give a brief introduction into artificial intelligence and more specifically machine learning (ML). We briefly summarise the history of artificial intelligence and machine learning, introduce the concepts of supervised learning, unsupervised learning, and reinforcement learning as the three main types of ML algorithms and discuss how to measure the quality of machine learning algorithms. For the interested reader, the appendix of the chapter includes a brief description of artificial neural networks and machine learning metrics.

1 Introduction

Artificial intelligence (AI) has become an important topic in research as well as industry. As stated in the Artificial Intelligence Index Report 2019, the number of peer reviewed publications on AI increased by more than 300% between 1998 and 2018 (Perrault et al., 2019). In the same report, large companies were surveyed out of which 58% stated that they were using AI in at least one function or business unit in 2019 (Perrault et al., 2019).

Especially in practice, the question often arises what is actually meant by AI. This question is not so easy to answer, because there are many definitions. Finding a uniform definition is difficult for two reasons: firstly, because of the breadth of the field, and secondly, because even a definition of *intelligence* proves difficult. There is general agreement that AI is a sub-area of computer science that deals with the research and development of so-called “intelligent agents” (Franklin & Graesser,

M. Reuter-Oppermann · P. Buxmann (✉)

Information Systems | Software & Digital Business, Technical University of Darmstadt, Darmstadt, Germany

e-mail: melanie.reuter-oppermann@tu-darmstadt.de; peter.buxmann@tu-darmstadt.de

1997). This is characterized by their ability to solve problems independently, without human interaction (Carbonell et al., 1983). Examples for AI include reasoning, knowledge representation, planning, learning, natural language processing, perception, and the ability to move and manipulate objects.

It is important to differentiate between strong and weak AI: strong AI is generally understood to include all approaches that attempt to depict and imitate humans or the processes in the brain. Properties such as consciousness or empathy are also frequently mentioned as constitutive characteristics of such a strong AI (Pennachin & Goertzel, 2007; Searle, 1980). However, research is still far from this point and we are not aware of any research projects that have come close to implementing strong AI. In contrast, solutions that are now technically feasible and have been implemented in current software solutions can be classified as weak AI (or narrow AI). Weak AI aims at developing algorithms specifically for certain, delimited problems (Goertzel, 2010; Pennachin & Goertzel, 2007).

An essential requirement not only for strong, but also for many weak AI, is the ability to improve their performance over time (Faraj et al., 2018), as a process of optimization on the basis of experience and the adaptation to possibly changing environmental conditions. In recent years, AI has developed more strongly in the direction of machine learning (ML). According to Brynjolfsson and McAfee (2017) of MIT, this is the most important basic technology of our age.

Research on early approaches of ML has actually been going on since the 1940s. Nevertheless, many technologies only become established when the framework conditions are right—as it is currently the case with AI applications. In recent years, the general conditions for the application of ML approaches have improved dramatically. Some barriers have been removed and new conditions have been created:

Firstly, digitized data is now available in an unprecedented amount—both on the Internet and in companies. This data is the basis for the use of ML methods. In addition, platforms such as Kaggle, a platform for a data science community owned by Google, exist, which provide tailored data sets for a variety of AI applications.

Secondly, computing power and storage space are more cost-effective than ever before and can be easily obtained from cloud providers. It is possible that developments in quantum computing will further help computationally intensive AI applications to make a breakthrough in the future. However, in addition to the advantages, it should also be considered that the use of cloud services can become a real cost factor when dealing with large amounts of data.

Third, the performance of ML algorithms has improved in recent years which enables them to be used on large scale, complex problems and applications.

Fourthly, there are many toolkits and libraries available free of charge for the development of AI applications. Examples include Scikit-learn, Apache Spark MLlib, Keras, CNTK, PyTorch or TensorFlow (Buxmann & Schmidt, 2019). Most of these tools were published under an open source license and facilitate the development of ML applications enormously. For example, tools such as Tensorflow or Scikit-Learn can be used to easily integrate ML methods into software code. In addition, frameworks like RapidMiner exist which support the entire development

process, including tasks like modeling as well as the processing, cleansing, and visualization of data. In addition, the possibilities to use ML algorithms have also been simplified by the fact that providers such as Google, IBM, Microsoft, or SAP now offer AI services based on a pay-per-use payment model. This means that users can obtain services such as voice to text conversion or object recognition via a software-as-a-service model. Business models are thus developing around the use of ML, which will further promote its use and distribution in the future.

From a technical perspective, the term *machine learning* generally encompasses methods that use learning processes to identify relationships in existing data sets in order to make predictions based on them (Murphy, 2012). There are many different concepts of the term. Often the approach of Tom Mitchell is used, who defines the basic concept of ML as follows: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.” (Mitchell, 1997, p. 2). In other words, the ability of a machine or software to learn certain tasks is based on the fact that it is trained on the basis of experience, i.e. data, instead of trying to explicate knowledge into hard coded rules and algorithms. What sounds harmless is a paradigm shift. Let us consider the recognition of cats, dogs or other animals in pictures as an example. In order to teach the algorithm a distinction, the developer no longer explicitly states in the software code that a cat has, for example, four paws, two eyes, sharp claws and fur. Instead, the algorithm is trained with many different animal photos, which it uses to learn how the respective animals look and how they differ from other animals. Another example to illustrate the basic principle are audio systems where an algorithm is trained with audio data containing a certain word, e.g. *destination input* for the navigation system in a car. In this way, the algorithm learns what this word sounds like, even if it is pronounced differently by different people or if there are different background noises. This is remarkable for several reasons: We humans often know more than we can express. This makes it difficult for software developers or analysts to code or specify certain facts. One speaks of the so-called Polanyi paradox, named after the philosopher Michael Polanyi: “We know more than we can tell” (Polanyi, 1966). This principle can be illustrated well when looking at Fig. 1: We can immediately tell that the left image displays a sheepdog and the right one shows a mop. Explaining why the image falls into a certain category is not trivial, though. In fact, when asking a popular stock image database for similar images to the mop, the search results include the sheepdog image. This lets us assume that an underlying AI considers both images to be similar. This idea was popularised by Karen Zack on her Twitter account, where she published several image sets on different themes, e.g. “sheepdog or mop”, “Chihuahua or muffin” or “puppy or bagel” (Zack, 2016). Many ML based systems are excellent learners, exceeding the capabilities of humans in many tasks, for example when diagnosing diseases or detecting fraud (Fawcett & Provost, 1997; Litjens et al., 2017).



Fig. 1 Sheepdog or mop? © Getty Images

2 The History of Artificial Intelligence and Machine Learning

The *Summer Research Project on Artificial Intelligence*, which took place at Dartmouth College in Hanover, New Hampshire, in 1956, is considered to be the birth of AI. It was a six-week conference organized by John McCarthy, the inventor of the programming language LISP. Other prominent participants were the AI researcher Marvin Minsky (1927–2016), the information theorist Claude Shannon (1916–2001), the cognitive psychologist Alan Newell (1927–1992) and the later Nobel Prize winner in economics Herbert Simon (1916–2001). The participants shared the view that intelligence can also be created outside the human brain. However, they disagreed on the way to achieve this, and the term *artificial intelligence* proposed by McCarthy remained controversial then—as it does today (Manhart, 2017).

Following this conference, AI research received a major boost as computers became faster and cheaper and the capacity to store data increased. Progress was also made in the field of Artificial Neural Networks (ANN). Demonstrators, such as the General Problem Solver developed by Newell and Simon or Joseph Weizenbaum’s program ELIZA already demonstrated the potential of AI algorithms.

However, these initial successes led to misjudgments and exaggerations. For example, Marvin Minsky told *Life Magazine* in 1970 “in 3 to 8 years we will have a machine with the general intelligence of an average human being”. Herbert Simon, who predicted in 1957 that within the next 10 years a computer would prove an important mathematical theorem, was subject to a similarly optimistic misjudgment (Newell & Simon, 1958). As a result, many expectations were not fulfilled at first, partly due to insufficient computing power. The period from 1965 to about 1975 is therefore often referred to as AI winter (Manhart, 2017).

In the 1980s, the focus was particularly on the development of so-called expert systems, led by Edward Feigenbaum, a computer science professor at Stanford University. The principle of expert systems is essentially based on a definition of

rules that explicitly formalize knowledge and the development of a knowledge base for a thematically clearly defined problem. The MYCIN system, which was used to support diagnosis and therapy decisions in blood infection diseases and meningitis, became particularly well known (Shortlife et al., 1975). Intensive research was also conducted on expert systems for operational applications (Mertens, 1985). Ultimately, however, these systems were not able to prevail, as rules were too rigid and the systems had a limited learning capacity.

Also, at the beginning of the 1980s, Japan set a clear signal in the direction of AI research with the so-called *Fifth Generation Project*, in which 400 million dollars were invested. The researchers' goals were primarily practical applications of AI. For implementation, they did not favor LISP, which was widely used in the United States, but tended towards the PROLOG language developed in Europe in the 1970s (Odagiri et al., 1997).

In 1990, Marvin Minsky initiated distributed AI as another new approach. This formed the basis of the so-called agent technology, which can be used for simulation-based analysis in various areas of investigation (Chaib-Draa et al., 1992). Also in the 1990s, great progress was made in the field of robotics. One competition with high publicity value is the RoboCup, in which scientists and students from all over the world let their robot teams compete against each other in soccer (Hess et al., 2014). This phase also saw the development of complex algorithms in the field of ANN (Nilsson, 2014; Russell & Norvig, 2010).

In 1997, the competition between IBM's Deep Blue and the then world chess champion Garry Kasparov caused a great public stir. Deep Blue narrowly won the duel with a score of 3.5:2.5, which was partly interpreted in the media as a victory of the computer over humankind. However, critics noted that Deep Blue was not really an intelligent system. Rather, the system simply used *brute force*, i.e. it simply calculated the consequences of all (halfway plausible) moves with high computing power. In fact, Deep Blue used heuristic algorithms that enable an intelligent search (Korf, 1997). Korf states that "if any technique deserves to be called AI, this one does."

3 Machine Learning Algorithms

When talking about ML algorithms, basically three types can be distinguished (Marsland, 2014; Murphy, 2012; Russell & Norvig, 2010): (1) supervised learning, (2) unsupervised learning, and (3) reinforcement learning.

Algorithms falling into the first category are trained with a lot of *labeled* data, i.e. input-output pairs, in order to learn a function that maps an input to an output. For example, an algorithm can be trained with several thousand cat and dog images. For each image the algorithm gets the information which animal species it is. That is, the input is labeled. In this way, supervised learning algorithms learn similarly to us humans. After the training, a test data set is used to make statements about the quality of the trained model. The actual learning process is thus based on a training data set,

while the evaluation of the trained model is carried out with a test data set (Marsland, 2014; Murphy, 2012; Russell & Norvig, 2010).

Unsupervised learning approaches try to find patterns in existing data. Let us take, for example, again a set of animal images. This time, the machine does not get the information which picture is which animal, but the algorithm has to find categories itself. A potential problem, but also a chance, is that the algorithm does the categorization on its own. The animal photos will not necessarily be categorized by animal species (dog or cat), but could alternatively, depending on the data situation, result in clusters by color (black, brown or white animals). Compression methods to filter out the least important components of the data and thus achieve a reduction in file size are another common application of unsupervised learning (Saul & Roweis, 2003). Other application areas for supervised learning include speech recognition, face recognition, fraud detection or traffic light management (Brynjolfsson & McAfee, 2017).

The third method category of ML is the so-called reinforcement learning. These methods are supposed to learn an optimal strategy for a given problem. The basis is an incentive or reward function that is maximized. The algorithm is not told which action is the best in which situation. Instead, it receives feedback on the selected action at certain points in time based on the incentive function—either a reward or a penalty. In this approach, the developer specifies the current state of the environment (e.g. the position in a chess game) and lists the possible action alternatives and environmental conditions (e.g. the possible moves based on the rules of the game). The algorithm must now find the moves that maximize its incentive function. In the case of chess, an incentive function would be specified in such a way that the objective is to win the game.

For a comprehensive and in-depth examination of ML methods see, for example, LeCun et al. (1998), Krizhevsky et al. (2012), Bishop (2006) or Hastie et al. (2009). The principle of supervised learning is most frequently applied today, as a great advantage of this principle is the variety of possible applications. In addition, numerous software tools are available, often on an open source basis, as for example Weka or scikit learn.

Another term that has been used very frequently in recent times is deep learning. This approach uses ANN as a basis. The basic idea behind the development of ANN is to simulate the (human) brain. In general, an ANN consists of nodes (neurons) and edges (synapses). Three types of neurons are distinguished, which are also called units. Ultimately, the acquired knowledge of an ANN is represented by the weights assigned to the edges between two nodes, which can be easily represented on the basis of matrices. Further explanation of how ANN work is given in the appendix (Section “Artificial Neural Networks”). These networks have a great advantage over previous generations of ML: With the help of multi-layered networks, they can learn interrelationships that remain hidden from simple ML algorithms. In addition, they benefit more from a larger amount of training data (Krizhevsky et al., 2012). While this might sound very complicated, it is fascinating how many software tools are available today on an open source basis, which can be used to develop AI-based algorithms in a quite simple way.

4 Measuring the Quality of Machine Learning Algorithms

ML approaches are suitable for a variety of application scenarios. In some cases, we are talking about a general-purpose technology (Brynjolfsson & McAfee, 2017). ML algorithms are not the *magic bag* for all more or less structured problems, even though they are or will be superior to humans in many areas, e.g. in some areas of medical diagnosis or object recognition. Rather, one has to be aware of existing limitations. The algorithms achieve very good results or decisions in many scenarios, although purely statistically speaking. However, this does not mean that they do not make wrong decisions. For example, a Google AI solution mistook a cat on a picture as guacamole (Anthalye et al., 2017). Another example is a turtle lying on its back, which the algorithm mistook for a rifle. In the case of the cat guacamole example, the algorithm was simply tricked. Knowing the parameters of the algorithm, the neural network can be outsmarted by a few added strokes or dots. These examples sound funny, but it becomes less fun when driving in an autonomous car and the algorithm mixes up traffic signs, for example.

Therefore, it is necessary to evaluate the quality of AI algorithms before their use. The basic principle shall be explained by the example of classification problems. These include object recognition on images (Dalal & Triggs, 2005) or medical diagnoses (Kononenko, 2001). To test how well solutions of ML approaches work for such problems, the so-called confusion matrix (CM) can be set up. This can be explained most easily using a simple example: A trained ML-based application is to be evaluated that uses health data to determine whether a patient has a disease or not. The application thus indicates which of the classes—*disease* versus *healthy*—a patient is classified into. In order to understand how well this classification works, *real* diagnoses are used as benchmarks that indicate whether a patient really has the disease or not. Thus, for each patient in the test data set, the actual class and the class predicted by the application are available. If the two patient classes are then compared, one of the following four cases is always present (Fawcett, 2006):

- True positive (RP): The patient has actually the disease and is classified as such by the application.
- True negative (RN): The patient has not actually the disease and is classified as such.
- False positive (FP): The patient has not actually the disease, but the application incorrectly classifies the patient as having it.
- False negative (FN): The patient actually has the disease, but the application incorrectly classifies the patient as not having it.

We then count how often these four cases occur in the classified test data to create the CM. Figure 2 shows the general structure of a CM. It consists of the dimensions *actual value* and *predicted value*, which divide the data into the four described cases. Each of the four resulting cells reflects the number of each case. The four values allow one to make a basic assessment of how well the application can correctly assign classes (Fawcett, 2006). For example, Fig. 3 shows that a corresponding

		actual value	
		1	0
predicted value	1	true positive	false positive
	0	false negative	true negative

Fig. 2 Structure of a confusion matrix

		actual value	
		disease	no disease
predicted value	disease	161	55
	no disease	18	302

Fig. 3 Example for a confusion matrix

model has classified relatively many patients as having the disease who are actually not having it—161 patients were correctly classified as having it, while 55 were incorrectly classified as having it. However, it is also apparent that the model can classify patients not having the disease relatively well—302 patients were correctly classified as not having it while only 18 were incorrectly classified.

The CM can be used to gain insight into the distribution of actual and predicted values. Organizations can use the information not only to understand how many errors an application produces in total, but more importantly to control potential error types. When relevant, ML algorithms can be parameterized during the development to reduce one type of potential error, but this will usually lead to an increase in the other type of error. For example, if it is more important that few false positives are generated, the algorithm can be more focused on this type of defect, but this can lead to an increase in false negatives. In some war-or-peace situations (such as an early warning of a nuclear attack) false positives can lead to unnecessary reactions and can therefore have extremely bad consequences so that special efforts need to be taken to avoid the latter, which are the more difficult the less time there is for a possible reaction.

A CM does not yet represent a concrete target value on the basis of which an ML solution can be optimized. More concrete quality measures can be established based on the values of the CM (Fawcett, 2006; Powers, 2011). As already emphasized at the beginning, these quality measures only refer to classification problems for which they have a certain standard status. For other types of problems there are less clear measures, which are often defined for each use case. These include, for example, root mean square error in forecasting (Hyndman & Koehler, 2006), (dis)similarity measures in clustering (Pfitzner et al., 2009) or health scores in predictive maintenance

(Gouriveau et al., 2016). The interested reader can find an overview over different metrics for measuring the quality of ML approaches in the appendix.

With all the positive aspects of AI in mind, it is also necessary to be aware of potential weaknesses and risks. Especially in application areas such as healthcare or the military, users need strong trust in the AI, for example when suggesting diagnosis decisions that can have a huge impact on a patient's life. Military AI applications pose particular problems. For many users in practice, ML approaches are experienced as a *black box* leading to a potential lack of trust. As a consequence, researchers have started to work on explainable AI, in order to increase the transparency of ML approaches. Another issue is the importance of balanced, complete and representative training data. If a certain value for one feature is overrepresented, there is an increased probability that the trained model is biased towards this value. For example, if you train a classification model only with dog images, a cat image will probably be also categorized as a dog. While sometimes, this is not an issue, it is important to be aware of it, especially if fairness is important and a bias must be prevented, e.g. regarding gender or race.

5 Conclusions

AI is on its way to changing society and the economy sustainably. While the development of strong AI is still unrealistic, weak AI approaches have come a long way in the last few years.

For a variety of application areas algorithms are already available in various tools, many of which are published open source. Therefore, AI is becoming an integral part of the business models of many providers. For many application areas, especially those that have not used AI before, a substantial *AI revolution* can be expected that can open unforeseen possibilities.

When developing individual AI applications, a major challenge is the availability and quality of data, especially training data. Often, data is actually the bottleneck nowadays, not algorithms or computation power. In order to successfully implement and employ AI applications, a new data culture is required in companies where, for example, silo mentality must be abandoned. Overall, data is increasingly becoming a competitive factor for companies—not only as the basis for AI applications, but also as the basis for a variety of strategic, tactical and operational decisions. The ICD, the premier global market intelligence firm, has predicted that in 2025 49% of the world's data will be stored in public cloud environments (Reinsel et al., 2018).

Future research and developments will also have to focus on data privacy and ethics when using AI in practice, e.g. when using patient data in healthcare applications, as well as with explainability and transparency of algorithms for users to understand and trust them.

Appendix

Artificial Neural Networks

The basic idea behind the development of ANN is to simulate the (human) brain. In general, an ANN consists of nodes (neurons) and edges (synapses). As the following figure shows, three types of neurons are distinguished, which are also called units (Goodfellow et al., 2016; Rey & Wender, 2018):

- Input units receive the input data, for example pixels in an image recognition algorithm or blood values when diagnosing diseases. Input units are denoted by x in Fig. 6.
- Hidden units are located between input and output units and thus represent the inner layers of an ANN. They can be arranged in several layers one after the other and are denoted by $h_1 \dots h_n$ in Fig. 4.

Output units contain the output data, for example a classification *dog* or *cat* in an algorithm for the recognition of animals. These are marked with y in Fig. 4.

A simple neural network contains only one hidden layer and is often already sufficient for many applications. Deep neural networks have multiple hidden layers, while the necessary or *good* number of layers and neurons depends on the individual application.

As the figure shows, the neurons are connected by edges, expressed as arrows. If we denote two neurons with i and j respectively, w_{ij} expresses the weight along the edge between i and j (Fig. 5).

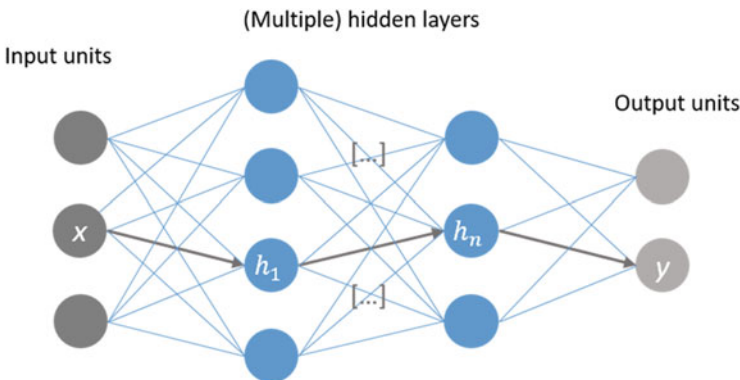


Fig. 4 Example for an artificial neural network

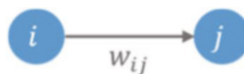


Fig. 5 Two neurons i and j and their respective weights w_{ij}

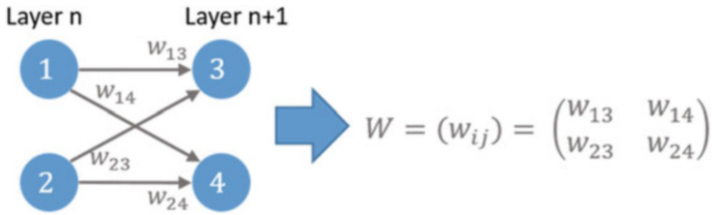


Fig. 6 Representation of the weights

Ultimately, the acquired knowledge of an ANN is represented by these weights, which can be easily represented on the basis of matrices (Fig. 6).

The input that one neuron receives from others depends on the output of the sending neuron(s) and the weights along the edges. If $Output_i$ denotes the activity level of a sending neuron i , then the input that a neuron j receives can be expressed as the sum over the weighted outputs of the neurons feeding it, adjusted with a bias offset value b_j , as in the following equation.

$$Input_j = \sum_i (Output_i * w_{ij}) + b_j$$

The output of a neuron is based on the input and an activation function. Various function types are conceivable for this activation function a —in the simplest case it is linear.

$$Output_i = a(Input_i)$$

The weights represent the knowledge of the ANN. These weights are modified based on learning rules. For example, when applying a supervised learning algorithm, the weights are modified or adjusted based on the training data. The most common procedure today is probably the so-called backpropagation method. Put simply, it works in such a way that errors in the initial layer are proportionately attributed to the error contributions of the hidden units involved and the weights are iteratively adjusted (Rumelhart et al., 1986).

Machine Learning Metrics

Different metrics exist to measure the quality of machine learning approaches, often depending on the type of approach that is used, e.g. classification, regression or deep learning. Note that a metric is different from a loss function. A loss function maps one or several variables to a real number and is often used as an objective function in mathematical optimization, for example. While metrics are usually used to measure

the performance of an approach, a loss function is used to train a machine learning approach.

Classification Metrics

For classification problems several metrics exist, including accuracy, precision, recall and the F1 score. They can all be computed based on the CM (see Fig. 3).

Classification Accuracy

Classification accuracy is computed as the ratio of the number of correct predictions to the total number of input samples. While it is a simple metric, it is problematic when the costs of one type of misclassification are very high. If a patient is wrongly classified as non-cancerous, for example, it can have fatal consequences.

In general, accuracy can be computed as:

$$accuracy = \frac{\textit{number of correct predictions}}{\textit{total number of predictions}}$$

With respect to the CM, accuracy can be computed by taking the values on the *main diagonal*:

$$accuracy = \frac{\textit{true positives} + \textit{true negatives}}{\textit{total number of predictions}}$$

Detection Rate

The detection rate gives the percentage of correctly predicted trues (or 1 s) with respect to the total number of predictions:

$$detection\ rate = \frac{\textit{true positives}}{\textit{total number of predictions}}$$

Precision

The precision or the positive predicted value gives the percentage of correctly predicted 1 s with respect to all predicted 1 s:

$$\mathit{precision} = \frac{\mathit{true\ positives}}{\mathit{true\ positives} + \mathit{false\ positives}}$$

Recall

A recall score measures the percentage of correctly predicted 1 s with respect to all actual 1 s. It is also called sensitivity or true positive rate:

$$\mathit{recall} = \frac{\mathit{true\ positives}}{\mathit{true\ positives} + \mathit{false\ negatives}}$$

Specificity

The specificity is also called the true negative rate. It determines the percentage of all 0 s that were correctly predicted:

$$\mathit{specificity} = \frac{\mathit{true\ negatives}}{\mathit{false\ positives} + \mathit{true\ negatives}}$$

Balanced Accuracy

The balanced accuracy is computed as the mean of recall and specificity and therefore balances the percentages of correctly predicted 1 s and 0 s:

$$\mathit{balanced\ accuracy} = \frac{\mathit{recall} + \mathit{specificity}}{2}$$

F1 Score

The F scores combine the precision and recall metrics. In general, an F score for a value β can be computed as:

$$F_{\beta} = (1 + \beta^2) \cdot \frac{\mathit{precision} \cdot \mathit{recall}}{\beta^2 \cdot \mathit{precision} + \mathit{recall}}$$

In the special case for $\beta = \mathbf{1}$, the F1 score is the harmonic mean between precision and recall. The range for the F1 score is $[\mathbf{0}, \mathbf{1}]$. The greater the F1 score, the better the performance of the model. F1 can be computed as:

$$F_1 = 2 \cdot \frac{\textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}}$$

Regression Metrics

Typical regression metrics are the mean absolute error and the mean squared error.

Mean Absolute Error

The mean absolute error is equal to the average of the absolute differences between the original values v_i and the predicted values w_i . It expresses how far the predictions were from the actual values. However, it does not give the direction of the error, i.e. whether data was over or under predicted. With N denoting the number of values, it can be computed as:

$$\textit{mean absolute error} = \frac{1}{N} \sum_{i=1}^N |v_i - w_i|$$

Mean Squared Error

The mean squared error and the mean absolute error are comparably similar. The only difference is that the mean squared error uses the average of the squares of difference between the original and the predicted values. Due to taking the square of the error, larger errors become more dominant compared to smaller errors. Therefore, when using the mean squared error, the focus is on larger errors:

$$\textit{mean squared error} = \frac{1}{N} \sum_{i=1}^N (v_i - w_i)^2$$

The root mean squared error takes the square root of the average of the squares of difference between the original and the predicted values and is therefore also sensitive to outliers.

References

- Anthalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2017). *Fooling neural networks in the physical world with 3D adversarial objects*. <https://www.labsix.org/physical-objects-that-fool-neural-nets/>
- Bishop, C. (2006). *Pattern recognition and machine learning*. Springer-Verlag New York.
- Brynjolfsson, E., & McAfee, A. (2017). The business of artificial intelligence. *Harvard Business Review*. <https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence>
- Buxmann, P., & Schmidt, H. (2019). *Künstliche Intelligenz*. Springer.
- Carbonell, J. G., Boggs, W. M., Mauldin, M. L., & Anick, P. G. (1983). The XCALIBUR project: A natural language interface to expert systems. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI'83)*, Morgan Kaufmann Publishers Inc., 653–656.
- Chaib-Draa, B., Moulin, B., Mandiau, R., & Millot, P. (1992). Trends in distributed artificial intelligence. *Artificial Intelligence Review*, 6(1), 35–66. <https://doi.org/10.1007/BF00155579>
- Dalal, N., & Triggs, B. (2005). *Histograms of oriented gradients for human detection*. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), 1, IEEE, 886–889. <https://doi.org/10.1109/CVPR.2005.177>.
- Dartmouth College. (1956). Summer research project on artificial intelligence. In *Volunteer officer experience (VOX) conference*. USA.
- Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1), 62–70. <https://doi.org/10.1016/j.infoandorg.2018.02.005>
- Fawcett, T., & Provorst, F. (1997). Adaptive Fraud Detection. *Data Mining and Knowledge Discovery*, 1(3), 291–316. <https://doi.org/10.1023/A:1009700419189>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Franklin, S., & Graesser, A. C. (1997). *Intelligent agents III. Lecture notes on artificial intelligence* (pp. 21–35). Springer-Verlag.
- Goertzel, B. (2010). Toward a formal characterization of real-world general intelligence. *Proceedings of the 3rd Conference on Artificial General Intelligence (AGI)*. Atlantis Press, 74–79. <https://doi.org/10.2991/agi.2010.17>.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Gouriveau, R., Medjaher, K., & Zerhouni, N. (2016). *From prognostics and health systems management to predictive maintenance I: Monitoring and prognostics*. Wiley.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer-Verlag New York.
- Hess, T., Legner, C., Esswein, W., Maaß, W., Matt, C., Österle, H., Schlieter, H., Richter, P., & Zarnekow, R. (2014). Digital life as a topic of business and information systems engineering? *Business & Information Systems Engineering*, 6(4), 247–253. <https://doi.org/10.1007/s12599-014-0332-6>
- Hyndman, R., & Koehler, A. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1), 89–109. [https://doi.org/10.1016/S0933-3657\(01\)00077-X](https://doi.org/10.1016/S0933-3657(01)00077-X)
- Korf, R. E. (1997). Does deep blue use artificial intelligence? *ICGA Journal*, 20(4), 243–245.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS'12)*, Curran Associates Inc (pp. 1097–1105). <https://doi.org/10.1145/3065386>
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. <https://doi.org/10.1109/5.726791>

- Litjens, G., Kooi, T., Ehteshami Bejnordi, B., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. <https://doi.org/10.1016/j.media.2017.07.005>
- Manhart, K. (2017). *Eine kleine Geschichte der Künstlichen Intelligenz*. <http://www.cowo.de/a/3330537>
- Marsland, S. (2014). *Machine learning: An algorithmic perspective*. Taylor & Francis, Inc.
- Mertens, P. (1985). Künstliche Intelligenz in der Betriebswirtschaft. In D. Ohse, A. C. Esprester, H.-U. Küpper, P. Stähly, & H. Steckhan (Eds.), *DGOR. Operations research proceedings* (pp. 285–292). Springer. https://doi.org/10.1007/978-3-642-70457-4_71
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.
- Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. The MIT Press.
- Newell, A., & Simon, H. (1958). Heuristic problem solving: The next advance in operations research. *Operations Research*, 6(1).
- Nilsson, N. J. (2014). *Principles of artificial intelligence*. Tioga Press.
- Odagiri, H., Nakamura, Y., & Shibuya, M. (1997). Research consortia as a vehicle for basic research: The case of a fifth generation computer project in Japan. *Research Policy*, 26(2), 191–207.
- Pennachin, C., & Goertzel, B. (2007). Contemporary approaches to artificial general intelligence. In B. Goertzel & C. Pennachin (Eds.), *Artificial General Intelligence* (pp. 1–30). Springer. https://doi.org/10.1007/978-3-540-68677-4_1
- Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Mishra, S., & Niebles, J. C. (2019). *The AI index 2019 annual report*. AI Index Steering Committee, Human-Centered AI Institute, Stanford University.
- Pfitzer, D., Leibbrandt, R., & Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, 19(3), 361–394. <https://doi.org/10.1007/s10115-008-0150-6>
- Polanyi, M. (1966). *The tacit dimension*. Peter Smith.
- Powers, D. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63. <https://doi.org/10.48550/arXiv.2010.16061>
- Reinsel, D., Gantz, J., & Rydning, J. (2018). *The digitization of the world: From edge to core*. IDC White Paper. <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>
- Rey, G. D., & Wender, K. F. (2018). Neuronale Netze. In *Eine Einführung in die Grundlagen, Anwendungen und Datenauswertung*. Hogrefe Verlag GmbH & Co. KG.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533–536. <https://doi.org/10.1038/323533a0>
- Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Prentice Hall International, Inc.
- Saul, L. K., & Roweis, S. T. (2003). Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4, 119–155.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457. <https://doi.org/10.1017/S0140525X00005756>
- Shortliffe, E. H., Davis, R., Axline, S. G., Buchanan, B. G., Green, C. C., & Cohen, S. N. (1975). Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research*, 8, 303–320.
- Zack, K. (2016). *Sheepdog or mop?* URL: <https://twitter.com/teenybiscuit/status/707670947830968320/photo/1>: Karen Zack via Twitter.

The Military Rationale for AI



Frank Sauer 

● آقای هوش مصنوعی ●

🕌 رسانه هوش مصنوعی دانشگاه تهران 🕌

@MrArtificialintelligence

Abstract This chapter introduces artificial intelligence (AI) and machine learning as major enablers of military innovation, especially regarding autonomy in weapons systems. It discusses the potential of AI where sensing, decision-making and acting are concerned. It also sheds light on the risks involved and questions claims about the effectiveness, reliability and trustworthiness of AI in military settings.

1 Introduction

“Artificial intelligence” (AI) is an umbrella term for a variety of different computing techniques and procedures that automate tasks which previously required the application of human intelligence. The goalposts of what is considered artificially intelligent are constantly moving—what was once considered AI (such as computers playing chess) is regarded as just another piece of software today. Despite its fuzziness, the term AI is unfortunately used ubiquitously. Machine learning (ML), more particularly the “deep learning” variant using artificial neural networks, is a technique within the field of AI that has been responsible for most of the progress that AI has made in the commercial sector over the last decade.

AI and ML are still poorly understood in that they are simultaneously over- and underestimated. They are overestimated because the terms “intelligence” and “learning” evoke the wrong associations in many observers, namely associations with human learning and human intelligence—both of which differ significantly from how ML-based AI works and what this technology is currently capable of. After all, ML-based artificial “intelligence” is limited to extremely narrow tasks. It is, as Gary Marcus (2018a, b) famously put it, “greedy” (hungry for immense amounts of data), “brittle” (failing spectacularly when confronted with tasks differing slightly from those it was trained and optimized for) and “opaque” (prone to inexplicable errors, making it a black box impossible to debug). In other words, even the currently most

F. Sauer (✉)
München, Germany
e-mail: frank.sauer@unibw.de

powerful ML-based AI is neither comparable to the immediate one-shot learning humans are capable of nor to the level of understanding and the flexible, generalized skills and problem-solving competences that come with human intelligence. At the same time, the implications of AI remain underestimated because prematurely deployed AI applications create serious risks.

AI is one of the major driving forces of the currently emerging fourth industrial revolution in which technologies in the digital, physical and biological spheres are generating new types of innovation. Almost all aspects of life are already or will eventually be touched by it—from the way humans communicate to the way they conduct commerce all the way to the instruments and institutions with which they govern themselves. Consequently, when asked in private about the rationale for employing AI in the armed forces, more than a few military officials around the globe phrase their answers something like this: “We have to do it because everyone else is doing it.”

In other words, first, a full-fledged “civil-military-fusion” race to insert more commercial AI into the military in general is already underway and, second, an unregulated dynamic involving greater autonomy, especially of weapons, is evolving.

Section “AI and the Military” will first sketch the political frame of reference for this race, followed by an overview of some of the areas where military AI currently being discussed is being deployed. Section “Weapon Autonomy and “Fighting at Machine Speed”” will then focus on the incentive for using AI to increase weapon autonomy and increase operational speed. Section “Conclusion: The Hype Is Real—And So Are the Risks” will offer critical concluding thoughts.

2 AI and the Military

AI allows for an increase in automation, that is, for machines to perform a variety of relatively complex tasks with minimal or no human assistance or supervision. Just as militaries were keen to adopt the steam engine, electricity or electronics in the past, they hope nowadays to benefit from this new innovation as well (Horowitz et al., 2018, pp. 3–5; DSB, 2016, pp. 6–11).

The primary frame of reference for this competitive process is the great power rivalry between the US, China and, although to a somewhat lesser extent, Russia (Sauer, 2019; Kania, 2020; Kühne, 2020, pp. 12–26). Within that geopolitical context, data is often described as the oil of the twenty-first century. This popular analogy is flawed (not least because data, unlike oil, is not a finite resource), but it nevertheless draws attention to a connection that is decisive for current great power dynamics in general (Horowitz, 2018) and automation in warfare in particular: Data—the capabilities and possibilities for collecting it, the capacity for processing it by means of AI, and thus the opportunity for converting it into military (and economic) power—will be of crucial importance to the world order in the age of digitization.

In the US, the Pentagon is seeking closer ties to technology companies in Silicon Valley and has declared the military use of big data and AI technologies a vital element in its overall strategy to retain conventional superiority—in what has come to be known as the “Third Offset Strategy” (Hagel, 2014)—and develop a generally more “lethal” military force. China stated in 2017 that its official goal was to achieve global leadership in AI innovation by 2030 (Kania, 2017, p. 4). Beijing is working on civil-military integration as well. The country also intends to make military use of the results of its breathtakingly fast race to catch up commercially in what has come to be known as the “intelligentisation” of warfare (Kania, 2017, 2019, 2020; Saalman, 2019). Meanwhile, Russian President Vladimir Putin (quoted in Vincent, 2017) famously put it as follows: “Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world.”

According to the Future of Life Institute’s global AI policy database (FLI, 2020), 36 countries around the world have issued national AI strategies—among them the US, China, Russia, Canada, India, Brazil, Japan, Australia and various European countries such as France, Germany, Italy, Poland, Sweden and Norway (Franke, 2019; Franke & Sartori, 2019). Regarding military AI in particular, the US has commissioned a series of studies and white papers to guide the policy-making process (DIB, 2019; DSB, 2016); similarly, France (2019, p. 3) has published an AI task force report which serves as the artificial intelligence strategy of France’s Armed Forces Ministry. With Department of Defense Directive 3000.09, the US remains the only country to have issued an explicit national doctrine for using AI-enabled autonomy in weapons systems (DoD, 2017 [2012]).

Generally speaking, the use of AI will affect the military—regarding both conventional and nuclear forces—wherever sensing, decision-making and taking action are involved.

AI is envisaged as improving the collection and subsequent analysis of data, enabling better human decision-making and improving command and control. Computer vision provides one example of this. It is already being used to reduce the workload of human analysts by sifting through huge amounts of visual data such as, for instance, video streams from drones, and thus enhancing intelligence, surveillance and reconnaissance (ISR) capabilities. A future step currently envisaged is a “common operating picture” to replace the multitude of different information sources gathered from various platforms and in various formats (Sayler & Hoadley, 2019, p. 12). The resulting information dominance and greater “battlespace awareness” (DoD, 2018, p. 18), potentially with even more granularity through the networking of small, distributed, persistent sensors systems, promises to fulfill the old hope of lifting the fog of war and, in this manner, also reducing the likelihood of receiving friendly fire—a major incentive to casualty-averse modern militaries (Schörnig & Lembcke, 2006). Targeting, too, is improved when visual (potentially combined with other kinds) data can be automatically analyzed, and potential targets highlighted on a screen for human operators. The US Army’s Advanced Targeting and Lethality Automated System (ATLAS) system for ground combat vehicles, for

example, supposedly makes it possible to engage targets three times faster than the comparable manual process (Tucker, 2019). Increased precision during an engagement *can* result from this increase in targeting speed. It is important, however, not to conflate automation, precision and discrimination in this context. Automation can increase a weapon systems' precision-strike capability which in turn allows the system to be used in a manner that discriminates better between legitimate military targets and civilians, thus increasing compliance with international humanitarian law (IHL). A good hypothetical example is a cruise missile autonomously aborting its attack run due to the unexpected presence of civilians in the vicinity of the target. However, an automated weapon could also be used—unlawfully—to specifically attack illegitimate targets with a high degree of precision. A swarm of small anti-personnel drones targeting and killing members of a specific group only, identified for instance by superficial features such as skin color, would be a hypothetical example in this case. In other words, the functionality of the weapons system, its capabilities regarding accuracy and error probabilities when attacking a target, and the way such a system is deployed on the battlefield must all be viewed separately when assessing the interplay of automation and IHL compliance.

AI-enabled handling of big data is also highly relevant to signals intelligence (SIGINT). It permits analysis of communications online for early detection and warning purposes on an unprecedented scale. Active operations are also enabled, of course. Information operations deploying artificially generated videos, photos or even just texts (for example by using OpenAI's powerful GPT-3 language generator) can be used to spread false information or influence public discourse—which is why they are currently also a cause for great concern due to their potentially destabilizing effects when they spread virally through social media (Sayler & Hoadley, 2019, pp. 11–12; Williams & Drew, 2020; Hersman, 2020).

In nuclear early warning, AI is conceptualized as helping with multisensor data fusion and analysis: “[S]tates could decide to automate additional components of early warning because autonomous systems can detect patterns and changes in patterns faster than humans. This could have potential benefits for nuclear security and stability, because well-functioning algorithms could give decision makers more time in a complex environment” (Horowitz, 2019a, p. 80). Similarly, pattern recognition could also help to strengthen parts of the nuclear enterprise against cyberattacks. Military planners anticipate that AI will improve systems monitoring and detection of anomalous behavior as well as autonomously generating patches of software vulnerabilities and mounting swift, automated responses (“hack backs”) (Sayler & Hoadley, 2019, p. 11).

Military logistics is another, comparably mundane field where AI is expected to have a major impact. One popular application of AI aims at improving maintenance for equipment and major weapons systems, allowing a shift from a one-size-fits-all approach to maintenance schedules tailored to every individual system (Sayler & Hoadley, 2019, pp. 10–11). Adaptive logistics such as autonomous vehicles carrying out just-in-time land-based or airborne deliveries of, for instance, fuel and ammunition to an area of operations, are another application, improving agility and supporting “dominant maneuver” on the battlefield (Allen & Chan, 2017; Sauer,

2018; Cuihong, 2019, p. 66; Kozyulin, 2019, p. 79; Davis, 2019; Boulanin et al., 2020; Kühne, 2020, pp. 35–44; DSB, 2016).

In addition to this wide range of applications, and as already alluded to in the discussion of AI-enabled targeting, militaries also intend to benefit from AI “at the sharp end,” that is, for immediate use in weapons systems.

3 Weapon Autonomy and “Fighting at Machine Speed”

The level of “autonomy”—as it has come to be called—in weapon systems is steadily increasing (Roff, 2016; Boulanin, 2016; Boulanin & Verbruggen, 2017; Scharre, 2018). The term “autonomy” is unfortunate because it is prone to be misleading for some people, resulting in misconceptions such as the anthropomorphization of weaponry and notions of humanoid robots going rogue (van Rompaey, 2019; Crotoof, 2018). Nonetheless, it is now commonly used.

Weapon autonomy is a military development of paramount importance; it has been described as “the third revolution in warfare, after gunpowder and nuclear arms” (FLI, 2015) and arguably represents the area where AI is used in the military with the most far-reaching consequences.

A functional understanding of weapon autonomy has found broad acceptance in the scholarly literature. It is also gaining acceptance in the diplomatic debate at the United Nations Convention on Certain Conventional Weapons (CCW) in Geneva, the epicenter of the global discussion of possible international regulation since 2014 (Rosert & Sauer, 2020). The functional understanding has emerged not least because the United States and the International Committee of the Red Cross (ICRC) have adopted it (DoD, 2017 [2012]; ICRC, 2016; Scharre, 2018).¹

From a functionalist point of view, the issue is best understood as one of autonomy *in* a weapons system, that is, of the machine rather than a human performing a certain function (or certain functions) during the system’s operation (Boulanin & Verbruggen, 2017). The upshot of focusing solely on who, human, or what, machine, performs specific functions is that clearly distinct definitions of autonomy and automaticity are not required—after all, there is no clear difference made between the two even in technical and engineering disciplines. In fact, the two terms can be—and often are—used interchangeably. In addition, a functionalist can remain agnostic regarding the underlying technology. In functional terms it makes little difference what exact process allows the weapon system to perform an operational function.

Every military operation concluding with an attack on a target can be systematized along the steps of a “kill chain” or “targeting cycle” (iPRAW, 2019). This

¹For a more detailed explanation of why adopting a functionalist rather than a categorical approach to conceptualizing weapon autonomy is indeed sensible for more than one reason, see Rosert and Sauer (2020).

includes finding, fixing, tracking, selecting and engaging the target (as well as assessing the effects afterwards). An autonomous weapon completes this entire targeting cycle—including the final stages of selecting and engaging the target with force—without human intervention (or even supervision, for that matter). The last two functions, which the ICRC calls “critical functions” (ICRC, 2016, p. 7), are focused on specifically because most of the political, legal and ethical risks currently being discussed with regard to weapon autonomy (Sauer et al., 2018) derive from handing control of the use of force from a human to a machine during the final stages of the targeting cycle. In fact, after announcing the ATLAS targeting system mentioned above, the US Army was forced to clarify that this system is not designed to delegate the firing decision to the machine (Tucker, 2019). At least for now, a human still pulls the trigger on the target the AI has—swifter than a human could—selected for engagement.

AI techniques such as ML are not necessarily required to give a weapon system autonomy (or, for that matter, automaticity) in the critical functions of target selection and engagement. The close-in weapon system Phalanx on navy ships or the Patriot missile-defense system, to name only two examples, have had this autonomous functionality for decades. Obviously AI is a very powerful enabler (Horowitz, 2018, p. 39). Thus, weapon autonomy is not new, but, for the first time, recent innovations in AI allow the development and use of weapon autonomy on a much larger scale. In effect, it is only recently that autonomous targeting has started to leave its former military niche applications, most prominently represented by terminal defense systems, such as the two mentioned above, and be adopted more broadly.

Autonomy renders constant communication and control links optional. The effect of making a weapons system autonomous, that is, delegating decision-making authority from a human operator to the unmanned system, is not restricted to enabling it—in contrast to a remotely controlled one—to continue operating even in environments where communication is degraded or denied (Scharre, 2020, pp. 105–110). More importantly, weapon autonomy means that the inevitable delay between a remote human operator’s command and the system’s response is eliminated. This results in far more rapid reaction times, generating a key tactical advantage over any adversarial system that is controlled remotely and thus necessarily slower. In fact, the prospects of gaining the upper hand by allowing for the completion of the targeting cycle at machine speed is, despite the accompanying ethical and legal misgivings, arguably the most important factor propelling current efforts to make weapons autonomous and remove human control entirely (Altmann & Sauer, 2017; Horowitz, 2019b, p. 769; Scharre, 2020, p. 62).

At the 2019 Dubai Airshow, Chief of Staff of the US Air Force Gen. David Goldfein presented a test exercise in which a US military satellite located an enemy navy vessel and instructed an airborne surveillance system to determine the exact location of the target. Coordinates were passed on to a command and control aircraft which selected a naval destroyer and tasked it with the attack. The only human involved in this entire process was located on the destroyer, and was responsible for releasing the anti-ship missile which had been algorithmically selected for the attack

(DefenseNews, 2019). Weapon autonomy in (widely distributed) systems is thus not a technology of the future. The human in examples such as ATLAS or the US Air Force's automated kill chain is present due to obligations derived from policy, not limits of technology.

4 Conclusion: The Hype Is Real—And So Are the Risks

In the introductory section, AI was described as simultaneously over- and underestimated—this is equally true for the civilian as well as the military world. Missy Cummings (2020) thus remains skeptical about the real transformative potential of AI in the military. She draws direct parallels to the commercial world, where current progress in AI, while delivering qualified successes, has at the same time been completely “overhyped”—all while remaining brittle and limited in its application, even in its most advanced implementations such as computer vision. She highlights especially the automotive industry's exaggerated promises for self-driving cars as a key example. “The inability of AI to handle uncertainty raises serious questions about how successful it will be in military settings,” she points out. What is more, the exaggerated expectations spilling over from the civilian into the military world, she suggests, could lead countries to merely pretend to have AI capabilities—a “fake it till you make it” phenomenon sometimes also called “fauxtimation” in the commercial world, when start-up companies pretend to have developed some new “AI” product while actually selling customers either ordinary software or cheap human labor disguised as results derived from AI. Within the geopolitical framework sketched out above, that is, the great power race for AI-enabled armed forces, the already observable practice of exaggeration and pretense perpetuates mutual misperceptions and unnecessarily inflates threat assessments (Cummings, 2020).

As a field of research, AI is many decades old. Deep learning has only very recently moved into the limelight, and many AI applications—civilian and military alike—still make use of “good old-fashioned” AI (GOFAI) techniques such as rules and decision trees, or a combination of GOFAI with ML. Nevertheless, the current dominance of “connectionism” over “symbolism” in the context of neural nets and deep learning is largely responsible for the exaggerated expectations AI is expected to fulfill. The idiosyncrasies and limitations of AI imply that military AI might in fact not deliver on many of the promises sketched out in sections “AI and the Military” and “Weapon Autonomy and “Fighting at Machine Speed””.

First, while it is true that modern militaries are forced to handle high volumes of data, it is important to consider this in a differentiated manner and ask what data is available when and where. Especially with regard to autonomous target recognition and subsequent engagement, the context of use is key. An anti-ship missile looking for specific silhouettes of navy ships to identify as targets, an application drawing on available data already in use and a seemingly easy target, provides one example. Another, much more difficult context of use would be the cruise missile mentioned

above that looks for civilians in targeted area to avoid collateral damage. Data availability aside, it is in fact unclear how this latter application could be realized given the technology of the present and the foreseeable future. After all, civilian-ness, albeit a cornerstone of IHL, is a fuzzy concept and is defined *ex negativo* (Rosert & Sauer, 2019, p. 370). Current machine learning systems, even with a high volume of training data available to them, lack the required capacity to understand social context at a very basic conceptual level. Not to mention the fact that, unlike in the commercial sector, training data for military applications is usually scarce.²

Second, both examples—the anti-ship missile looking for navy vessels and the cruise missiles trying to avoid civilians—fail to take account of the fact that deep learning-based AI applications such as computer vision are prone to error and manipulation. Self-driving cars go through stop signs minimally altered with reflective tape, and even the most advanced image recognition algorithms can be tricked into confusing turtles with rifles or school buses with snow plows (Athalye et al., 2018; Marcus, 2018a, c), in addition to the problems arising from biased datasets (Vincent, 2020; Saylor & Hoadley, 2019, pp. 29–31; Horowitz et al., 2020; Cummings, 2020). Clearly then, overreliance on AI targeting also opens up new options for adversaries clever enough to deceive the system. A handful of soldiers dressed and behaving like civilians would fool the hypothetical cruise missile, if it were ever developed and used in the field, into aborting its operation. A few retractable dummy structures and antennas deployed by ships under attack could deceive the “smart” anti-ship missile. The fact that very expensive AI-enabled weaponry could potentially be tricked quite easily and cheaply, a fact derivable from knowing the capabilities and limits of the underlying technology, is rarely discussed—because it runs counter to the dominant military AI narrative.

Consequently, while it is currently open to question whether AI will eventually achieve the levels of reliability and trustworthiness that would be required for many of the applications currently envisaged for military purposes, what is emerging with absolute certainty are the risks of a new arms race. One risk indicator is the ongoing arms dynamic with regard to weapon autonomy that fuels strategic instability and risks of inadvertent escalation (Altmann & Sauer, 2017; Scharre, 2020). Another indicator is the potential crossing of legal and moral boundaries: Weapon autonomy, recklessly and prematurely applied in an effort to retain the technological edge in the arms race, might end up failing to improve and in fact degrading compliance with IHL. At the same time, it risks infringing on the dignity of human beings—whether combatants or civilians—by reducing them to a mere data point and having

²Observers who expect the development of weapon autonomy to allow for an overall increase in IHL compliance thus generally argue in a two-step fashion. First, weapons selecting and engaging targets autonomously would be used only in circumstances where no civilians and civilian infrastructure are present, thus removing the necessity of even having to discriminate between lawful and unlawful targets. Later, in a second step, technology would have matured enough for the system to be able to perform IHL compliance at least at a human level on its own. See Schmitt and Thurnher (2013, pp. 246–248); Anderson and Waxman (2013, pp. 11–13); Anderson et al. (2014, pp. 405–406).

machines snuff out life without human consideration or accountability (Asaro, 2012; Sparrow, 2016; Amoroso & Tamburrini, 2017; Rosert & Sauer, 2019; Skerker et al., 2020).

In sum, a widespread misjudgment of AI's strengths and weaknesses is in large part responsible for making exaggerated claims for the introduction of AI in military applications. Accusations of sensationalism (Ford, 2020, pp. 1–2) made against those favoring regulation who put forward arguments such as the ones presented above, some of whom make use of “killer robot” rhetoric (KRC, 2019), will remain ineffective as long as militaries are deceived by exaggerated claims and continue to label programs for AI-enabled drones “Skyborg” (Gunzinger & Autenreid, 2020). Instead of acknowledging the limits of excessive claims regarding technology and treating the militarizing of AI with the requisite cautiousness, prudence, and international regulatory safety measures, now a blind race is underway in which everything in the military nowadays is about to become “enhanced” or “powered” by AI in some shape or form. But without precautions and safeguards against the accompanying risks, the long-term drawbacks are certain to outweigh the short-term benefits.

Further Reading

- Defense Innovation Board. (2019). AI principles—recommendations on the ethical use of artificial intelligence by the department of defense.
- Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. W W Norton & C. Boulanin, V., Davison, N., Goussac, N., & Carlsson, M. P. (2020). *Limits on autonomy in weapon Systems: Identifying practical elements of human control*. SIPRI/ICRC report.

References

- Allen, G., & Chan, T. (2017). *Artificial intelligence and national security*. Accessed July 24, 2017, from <http://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf>
- Altmann, J., & Sauer, F. (2017). Autonomous weapon systems and strategic stability. *Survival*, 59(5), 117–142. <https://doi.org/10.1080/00396338.2017.1375263>
- Amoroso, D., & Tamburrini, G. (2017). The ethical and legal case against autonomy in weapons systems. *Global Jurist*, 18(1). <https://doi.org/10.1515/gj-2017-0012>
- Anderson, K., & Waxman, D. (2013). Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can. Jean Perkins Task Force on National Security and Law Essay Series, from https://scholarship.law.columbia.edu/cgi/viewcontent.cgi?article=2804&context=faculty_scholarship
- Anderson, K., Reisner, D., & Waxman, M. (2014). Adapting the law of armed conflict to autonomous weapon systems. *International Law Studies*, 90(386), 386–411.
- Asaro, P. (2012). On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross*, 94(886), 687–709.

- Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust Adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning PMLR 80* (pp. 284–293).
- Boulanin, V. (2016). *Mapping the development of autonomy in weapon systems: A primer on autonomy*. Accessed Jan 17, 2017, from <https://www.sipri.org/sites/default/files/Mapping-development-autonomy-in-weapon-systems.pdf>
- Boulanin, V., & Verbruggen, M. (2017). *Mapping the development of autonomy in weapon systems*. Accessed Nov 14, 2017, from https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_0.pdf
- Boulanin, V., Saalman, L., Topychkanov, P., Fei Su, & Carlsson, M. P. (2020). *Artificial intelligence, strategic stability and nuclear risk*. Accessed Jun 22, 2020, from https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf
- Crootof, R. (2018). Autonomous weapon systems and the limits of analogy. *Harvard National Security Journal*, 9(2), 51–83.
- Cuihong, C. (2019). The shaping of strategic stability by artificial intelligence. In L. Saalman (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk, vol. II, east Asian perspectives* (pp. 54–77).
- Cummings, M. (2020). *The AI that wasn't there: Global order and the (Mis)Perception of powerful AI: Texas national security review policy roundtable: Artificial intelligence and international security*. Accessed Jun 15, 2020, from <https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security/>
- Davis, Z. (2019). Artificial intelligence on the battlefield: Implications for deterrence and surprise. *PRISM*, 8(2), 115–131.
- DefenseNews. (2019). *Video: Here's how the US Air FORCE is automating the future kill chain. Chief of Staff of the U.S. Air Force Gen. David Goldfein Dubai Airshow 2019*. Accessed Jan 27, 2020, from <https://www.defensenews.com/video/2019/11/16/heres-how-the-us-air-force-is-automating-the-future-kill-chain-dubai-airshow-2019/>
- DIB. (2019). *Defense innovation board: AI principles—recommendations on the ethical use of artificial intelligence by the department of defense*. Accessed Jun 14, 2020, from https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF
- DoD. (2017 [2012]). *Directive 3000.09: Autonomy in weapon systems*. Accessed Dec 10, 2020, from https://fas.org/irp/doddir/dod/d3000_09.pdf
- DoD. (2018). *U.S. department of defense unmanned systems integrated roadmap 2017–2042*. Accessed Jun 13, 2020, from <https://assets.documentcloud.org/documents/4801652/UAS-2018-Roadmap-1.pdf>
- DSB. (2016). *Defense science board summer study on autonomy*. US department of defense. Accessed Dec 10, 2020, from <https://www.hsdl.org/?view&did=794641>
- FLI. (2015). *Autonomous weapons: An open letter from AI & robotics researchers*. Accessed Aug 31, 2015, from <https://futureoflife.org/open-letter-autonomous-weapons/>
- FLI. (2020). *National and international AI strategies: A global landscape*. Accessed July 25, 2020, from <https://futureoflife.org/ai-policy/>
- Ford, C. A. (2020). AI, human-machine interaction, and autonomous weapons: Thinking carefully about taking “killer robots” seriously. *Arms Control and International Security Papers*, 1(2).
- France. (2019). *French ministry of defense report of the AI task force: Artificial intelligence in support of defense*. Accessed July 25, 2020, from https://www.defense.gouv.fr/content/download/573877/9834690/Strat%C3%A9gie%20de%20I%27IA-UK_9%20I%202020.pdf
- Franke, U. E. (2019). *Harnessing artificial intelligence*. Accessed July 24, 2020, from https://www.ecfr.eu/publications/summary/harnessing_artificial_intelligence
- Franke, U. E., & Sartori, P. (2019). *Machine politics: Europe and the AI revolution*. Accessed July 24, 2020, from https://www.ecfr.eu/publications/summary/machine_politics_europe_and_the_ai_revolution

- Gunzinger, M., & Autenreid, L. (2020). *The promise of Skyborg*. Accessed Dec 10, 2020, from <https://www.airforcemag.com/article/the-promise-of-skyborg/>
- Hagel, C. (2014). *Reagan national defense forum keynote*. Accessed Nov 11, 2016, from <http://www.defense.gov/DesktopModules/ArticleCS/Print.aspx?PortalId=1&ModuleId=2575&Article=606635>
- Hersman, R. (2020). *Wormhole escalation in the new nuclear age*. Accessed July 25, 2020, from <https://tnsr.org/2020/07/wormhole-escalation-in-the-new-nuclear-age/>
- Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*, 1(3), 37–57. <https://doi.org/10.15781/T2639KP49>
- Horowitz, M. C. (2019a). Artificial intelligence and nuclear stability. In V. Boulanin (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk* (pp. 79–83).
- Horowitz, M. C. (2019b). When speed kills: Lethal autonomous weapon systems, deterrence and stability. *Journal of Strategic Studies*, 42(6), 764–788. <https://doi.org/10.1080/01402390.2019.1621174>
- Horowitz, M. C., Allenz, G. C., Kaniaz, E. B., & Scharrez, P. (2018). *Strategic competition in an era of artificial intelligence*. Accessed July 25, 2018, from https://s3.amazonaws.com/files.cnas.org/documents/CNAS-Strategic-Competition-in-an-Era-of-AI-July-2018_v2.pdf?mtime=20180716122000
- Horowitz, M. C., Kahn, L., & Ruhl, C. (2020). *Introduction: Artificial intelligence and international security: Texas national security review policy roundtable: Artificial intelligence and international security*. Accessed July 25, 2020, from <https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security/>
- iPRAW. (2019). *Focus on human control*. iPRAW Report No. 5. Accessed Feb 12, 2019, from https://www.ipraw.org/wp-content/uploads/2019/08/2019-08-09_iPRAW_HumanControl.pdf
- ICRC. (2016). *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*. <https://shop.icrc.org/download/ebook?sku=4283/002-ebook>. Accessed 11 April 2016.
- Kania, E. B. (2017). *Battlefield singularity: Artificial intelligence, military revolution, and China's future military power*. Accessed May 29, 2018, from <https://s3.amazonaws.com/files.cnas.org/documents/Battlefield-Singularity-November-2017.pdf?mtime=20171129235805>
- Kania, E. B. (2019). *In military-civil fusion, china is learning lessons from the united states and starting to innovate*. <https://thestrategybridge.org/the-bridge/2019/8/27/in-military-civil-fusion-china-is-learning-lessons-from-the-united-states-and-starting-to-innovate>
- Kania, E. B. (2020). *"AI weapons" in China's military innovation*. Accessed May 4, 2020, from https://www.brookings.edu/wp-content/uploads/2020/04/FP_20200427_ai_weapons_kania_v2.pdf
- Kozyulin, V. (2019). Regulatory frameworks for military artificial intelligence. In L. Saalman (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk, vol. II, east Asian perspectives* (pp. 78–85).
- KRC. (2019). *Global poll shows 61% oppose Killer Robots*. Accessed Jan 20, 2020, from <https://www.stopkillerrobots.org/2019/01/global-poll-61-oppose-killer-robots/>
- Kühne, S. (2020). *Das Versprechen von Künstlicher Intelligenz: Erste Ergebnisse einer Untersuchung zu Erwartungen an moderne Waffensysteme*. Accessed Jun 2, 2020, from https://ifsh.de/file/publication/Research_Report/003/20200525_IFSH_Research_Report_003_KI.pdf
- Marcus, G. (2018a). *Deep learning: A critical appraisal*. Accessed Feb 5, 2018, from <https://arxiv.org/ftp/arxiv/papers/1801/1801.00631.pdf>
- Marcus, G. (2018b). *Greedy, brittle, opaque, and shallow: The downsides to deep learning*. Accessed Dec 10, 2020, from <https://www.wired.com/story/greedy-brittle-opaque-and-shallow-the-downsides-to-deep-learning/>
- Marcus, G. (2018c). *The deepest problem with deep learning*. Accessed Dec 10, 2020, from <https://medium.com/@GaryMarcus/the-deepest-problem-with-deep-learning-91c5991f5695>

- Roff, H. M. (2016). *Weapon autonomy is rocketing*. Accessed May 10, 2016, from <http://foreignpolicy.com/2016/09/28/weapons-autonomy-is-rocketing/>
- Rosert, E., & Sauer, F. (2019). Prohibiting autonomous weapons: Put human dignity first. *Global Policy*, 94(1), 42. <https://doi.org/10.1111/1758-5899.12691>
- Rosert, E., & Sauer, F. (2020). How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies. *Contemporary Security Policy*, 1–26. <https://doi.org/10.1080/13523260.2020.1771508>
- Saalman, L. (2019). Exploring artificial intelligence and unmanned platforms in China. In L. Saalman (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk, vol. II, east Asian perspectives* (pp. 43–47).
- Sauer, F. (2018). *Artificial intelligence in the armed forces: On the need for regulation regarding autonomy in weapon systems: Security policy working Paper No. 26/2018*. https://www.baks.bund.de/sites/baks010/files/working_paper_2018_26.pdf
- Sauer, F. (2019). *Great powers and digitisation: What are the implications for world order?* Accessed Jan 14, 2020, from https://metis.sowi.unibw-muenchen.de/img/publications/08_10_2018_great_powers_and_digitisation.pdf
- Sauer, F., Amoroso, D., Sharkey, N., Suchman, L., & Tamburrini, G. (2018). *Autonomy in weapon systems: The military application of artificial intelligence as a litmus test for Germany's new foreign and security policy*. Accessed May 29, 2018, from https://www.boell.de/sites/default/files/boell_autonomy-in-weapon-systems_v04_kommentierbar_1.pdf
- Sayler, K. M., & Hoadley, D. S. (2019). *Artificial intelligence and national security*. CRS Report R45178. Accessed July 16, 2020, from <https://fas.org/sgp/crs/natsec/R45178.pdf>
- Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. W W NORTON & CO.
- Scharre, P. (2020). *Autonomous weapons and stability: Thesis submitted in fulfillment of the requirements for the degree of doctor of philosophy*. Accessed Jun 13, 2020, from https://kclpure.kcl.ac.uk/portal/files/129451536/2020_Scharre_Paul_1575997_thesis.pdf
- Schmitt, M. N., & Thumher, J. S. (2013). “Out of the loop”: Autonomous weapon systems and the law of armed conflict. *Harvard National Security Journal*, 4, 231–281.
- Schörnig, N., & Lembcke, A. C. (2006). The vision of war without casualties: On the use of casualty aversion in armament advertisements. *Journal of Conflict Resolution*, 50(2), 204–227. <https://doi.org/10.1177/0022002705284827>
- Skerker, M., Purves, D., & Jenkins, R. (2020). Autonomous weapons systems and the moral equality of combatants. *Ethics and Information Technology*, 3(6) online first.
- Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics & International Affairs*, 30(1), 93–116.
- Tucker, P. (2019). *US military changing ‘Killing Machine’ Robo-tank program after controversy*. Accessed July 24, 2020, from <https://www.defenseone.com/technology/2019/03/us-military-changing-killing-machine- robo-tank-program-after-controversy/155256/>
- van Rompaey, L. (2019). Shifting from autonomous weapons to military networks. *Journal of International Humanitarian Legal Studies*, 10(1), 111–128. <https://doi.org/10.1163/18781527-01001011>
- Vincent, J. (2017). *Putin says the nation that leads in AI ‘will be the ruler of the world’*. Accessed Feb 24, 2020, from <https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world>
- Vincent, J. (2020). *What a machine learning tool that turns Obama white can (and can't) tell us about AI bias*. Accessed July 25, 2020, from <https://www.theverge.com/21298762/face-depexelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias>
- Williams, H., & Drew, A. (2020). *Escalation by Tweet: Managing the new nuclear diplomacy*. Accessed July 17, 2020, from <https://www.kcl.ac.uk/csss/assets/10957%E2%80%A2twitterconflictreport-15july.pdf>

Military AI Applications: A Cross-Country Comparison of Emerging Capabilities



Sophie-Charlotte Fischer

● آقای هوش مصنوعی ●

رسانه هوش مصنوعی دانشگاه تهران

@MrArtificialintelligence

Abstract A new wave of progress in artificial intelligence (AI) is attracting considerable interest from militaries across the globe and a growing number of countries are already actively pursuing military AI capabilities. The purpose of this chapter is to provide an accessible review and comparison of the emerging military AI capabilities of four countries—the United States, China, France, and Israel. While only a preliminary assessment is possible at present, important clues can be derived from analyzing and comparing the kind of applications that states are investing in and the selected areas in which they are already deploying AI. Related to these aspects, the chapter also examines whether the four countries have introduced any restrictions for the military use of AI, so far.

1 Introduction

Militaries routinely acquire new technologies because they have the potential to enhance existing capabilities or to enable entirely new ones. Novel technologies can also have a significant impact on how wars are fought and alter the strategic environment in which military forces operate (James, 2013). Currently, a new wave of progress in artificial intelligence (AI)—“the ability of machines to fulfill tasks that normally require human intelligence” such as “recognizing patterns, learning from experience, drawing conclusions, making predictions, or taking action”—is attracting considerable interest from militaries across the globe (U.-S. Department of Defense, 2018, p. 5).

During his confirmation hearing in 2019, former US Secretary of Defense Mark Esper stated that “whoever masters it [AI] first will dominate [on the] battlefield for many, many, many years” (2019, p. 64). Esper’s statement reflects the extravagant promises associated with military AI applications: current assessments suggest that AI could significantly improve accuracy and speed in areas ranging from military

S.-C. Fischer (✉)
ETH Zurich, Zurich, Switzerland
e-mail: sophie.fischer@sipo.gess.ethz.ch

logistics to decision-making on the battlefield (Scharre, 2020). However, the continuing development of AI complicates assessment of its military effects and makes it difficult to separate fact from fiction. Moreover, safety and reliability considerations as well as ethical and legal concerns could restrict how militaries use AI (Horowitz, 2018).

Nonetheless, an increasing number of countries are already exploring military AI applications. While many of these efforts are still in their initial stages, I argue that important clues can be derived from analyzing and comparing how different countries are pursuing military AI such as the kind of applications they are investing in and the selected areas in which they are already deploying AI. This can amplify the nascent academic and policy discourse on how AI impacts the military and the broader issue of security. Related to these aspects is the question of whether countries have already put restrictions in place for the military use of AI. The purpose of this chapter is to provide an accessible review and comparison of the emerging AI capabilities of four countries—the United States, China, France and Israel—in the military realm.

2 Assessing Emerging Military AI Capabilities: A Framework for Analysis

How can we take stock of and compare the emerging military AI capabilities of different countries? The answer to this question is far from straightforward. The emerging nature of AI, its character as an enabling technology, and the lead of the commercial sector in AI development each make an assessment more complicated. In this section, I will identify three dimensions—the sources of military AI, areas of military AI application, and the risks associated with military AI applications—and base the subsequent analysis of the four countries on these dimensions.

2.1 The Sources of Military AI

During the Cold War, there were often trickle-down effects from military technology to the civilian realm (Alic et al., 1992). However, today, private-sector tech companies are leading the way in many high-tech areas, including AI (Villani, 2019; Laskai, 2018). Consequently, in order to bring AI into the military realm, militaries need to look to and increasingly collaborate with technology companies that are primarily focused on commercial markets.

However, to further complicate the assessment, a strong commercial AI sector is not necessarily equivalent to powerful military AI capabilities. This is not only due to the potential unwillingness of commercial companies to collaborate with militaries (Wakabayashi & Shane, 2018). Several military problems cannot be directly

solved by means of commercial AI applications. While some may be directly transferable “off-the-shelf,” others must undergo significant modification before they are viable for use by the military, a process which poses substantial integration challenges (Saylor, 2019). Moreover, for some military AI applications, commercial technology may be unsuitable and defense organizations or traditional defense contractors need to develop them. As a result, it is far from straightforward to infer that a strong commercial AI industry automatically translates into strong military AI capabilities.

Nevertheless, and despite this caveat, when evaluating what (potential) military AI capabilities countries have and because many of the underlying AI technologies are dual-use, looking at defense spending and the work of traditional defense contractors alone is not sufficient. AI investment in the civilian realm and militaries’ strategies to develop AI must also be taken into account. For these reasons, I will assess investments in both sectors and also examine whether and, if so, how militaries collaborate with civilian AI companies.

2.2 Areas of Military AI Applications

When looking at applications of AI in the military realm, it is essential to underline that AI is neither a weapon nor a discrete technology, but instead can be viewed as an enabling technology akin to electricity (Scharre, 2020). AI has a broad range of applications across many other technology fields in the civilian as well as the military realm (Horowitz, 2018). These properties of AI make the assessment of states’ military AI capabilities even more complicated, as they are difficult to grasp and cannot simply be quantified like nuclear warheads, for example. However, it is possible to investigate in what areas AI could be applied and assess states’ ongoing or planned activities.

Given the enabling character of AI, there is a seemingly infinite number of applications in the military realm (see also the text by Frank Sauer in this volume). At present, AI is expected to be particularly useful in intelligence, surveillance, reconnaissance (ISR) because it excels at analyzing and finding patterns in vast amounts of data (Franke, 2019). Moreover, AI has the potential to strengthen both offensive and defensive cyber capabilities. AI could also be deployed for information operations in order to either create content such as deep fakes or to identify forgeries generated using AI (Saylor, 2019). Weapon system command and control is another area where AI may prove useful. AI systems could gather and fuse information from all military domains and provide commanders with a range of actionable options based on their assessment (Horowitz et al., 2018). Similarly to the civilian realm, militaries could increasingly deploy semiautonomous and autonomous vehicles and use them to explore and operate in potentially hostile environments. Currently, the most controversial military AI application is in lethal autonomous weapons systems (LAWS), which denotes weapons system that could autonomously identify and engage targets (see also the text by Anja Dahlmann in

this volume). Last but not least, other more mundane but promising applications include the training of soldiers as well as military logistics, where AI could be used for tasks such as predictive maintenance of military aircraft (Sayler, 2019).

The different areas of application outlined above provide guidance for assessing where countries are already investing in AI, are already using it, or plan to do so in the future. It is of particular interest, whether the areas of AI applications that countries focus on are markedly similar or differ across countries, and if so how.

2.3 Risks Posed by Military AI Applications

While AI holds the promise of delivering significant military advantages across a number of application areas, its deployment can also introduce significant risks. Some of these risks are technological and operational. While the development of AI has advanced significantly over the last two decades, current AI systems are still brittle and often behave unpredictably. Other risks include, for example, adversarial attacks, data poisoning and reward hacking (Scharre, 2020). Moreover, the deployment of AI systems for military purposes also raises ethical, moral, and legal challenges, especially in relation to the possible use of LAWS. Since 2014, states have been discussing questions related to emerging technologies in the area of LAWS in the framework of the United Nations (UN) Convention on Certain Conventional Weapons (CCW).

In order to avoid some of these risks, states may want to constrain the development, production and/or use of AI, at least in certain areas of application or in certain contexts. Such regulations could impose important limitations on states' deployment of military AI capabilities. In the subsequent review, I will therefore also assess whether countries have already imposed limitations on the development and/or use of AI in the military realm and if so, what form these limitations take.

3 Assessing Military AI Capabilities: A Review of Four Countries

The objective of this section is to provide detailed insights into the emerging military AI capabilities of different countries. However, this level of detail means narrowing the focus to a small number of states. The choice of countries is not strictly representative, but also based on three factors: a clear intent to deploy AI in the military realm, the size and status of the domestic civilian AI industry, and the geostrategic context the states are situated in. The first two countries that I analyze are the US and China, which are considered "AI superpowers" based on the size and status of their civilian AI industries. Furthermore, the two countries are of particular interest, as the US retains the most powerful military globally, while China has

invested heavily in military modernization in recent decades and is regarded as the US' most likely contender for supremacy. Importantly, however, while most attention has, so far, focused on the US and China, AI could also empower other prosperous and technologically advanced states and affect their status in the international system (Barsade & Horowitz, 2018). I chose to study France and Israel as well because they are both classified as "second-tier" civilian AI powers, have significant military capabilities and represent two very different geostrategic contexts.

Based on the case selection criteria outlined above, I decided not to assess a number of other interesting countries in this chapter. Russia most notably is also actively pursuing military AI capabilities, but has a comparatively weak civilian-sector AI industry and clearly trails many other nations in terms of AI commercialization and cutting-edge research. However, although Russia is excluded from this study on these grounds, it could be a very interesting case for testing the link between the strength of states' civilian-sector AI industry and their military AI capabilities in future research. As some have argued, Russia was never a leader in the development of internet technology, but has become a highly disruptive force in cyberspace nevertheless (Allen, 2017; Sayler, 2020). This thus poses the question whether a similar pattern with regard to Russia's AI capabilities will be seen in the future.

3.1 The United States

Since the beginning of the Cold War, the US advantage in technology has underscored its military superiority. The Department of Defense's (DoD) First and Second Offset Strategy in the 1950s and 1980s respectively reflected the US military's focus on technological quality rather than quantity. However, as a consequence of especially China's military modernization, the relative technological advantage of the US has been eroding over the past two decades. The increasing interest of the US military in AI can be traced back to the US Third Offset Strategy that was presented by then Secretary of Defense Chuck Hagel in 2014 and is driven by both novel technological opportunities and broader geopolitical transformations (Hagel, 2014). AI is one of several technologies that the US is seeking to develop in order to maintain its technological edge and sustain a strategic advantage over competitors (Ellman et al., 2017).

The United States is regarded as the world's current leader in AI, because of its unrivaled lead in workforce talent, hardware and quality of publications and patents in the AI field (Ding, 2018; Breitinger et al., 2020). In 2019, former US President Trump signed an executive order on AI which laid the foundation for the US Federal Government's American AI Initiative (Saslow, 2020). The initiative has five key pillars: research and development (R&D), infrastructure, governance, workforce, and international engagement. However, already in 2018, the Pentagon specified its AI ambitions in the DoD's military AI strategy document and the separate US Airforce AI strategy annexed to it (U.S. Department of Defense, 2018).

3.1.1 Sources of Military AI

The implementation of these plans is backed by significant investments, although critics lament the fact that the current AI budget of the DoD is no more than a starting point. Defense-related spending on AI was around \$4 billion in 2020 (Harper, 2020). What is noteworthy, however, is that the US government allotted the major share of its overall spending on AI for 2020 to the military rather than the civilian sector (Hao, 2019). In order to streamline the development and adoption of AI, the Pentagon founded the Joint Artificial Intelligence Center (JAIC) in 2018 (Leung & Fischer, 2018). The military services, Defense Advanced Research Projects Agency (DARPA), and the Intelligence Advanced Research Projects Agency (IARPA) carry out AI R&D (Sayler, 2019).

Nonetheless, the DoD is acutely aware that the bulk of cutting-edge AI R&D is currently taking place in the private sector. The 2017 National Security Strategy, for example, highlighted the need to foster collaboration with private sector actors (The White House, 2017). To that end, the Pentagon is seeking to access commercial innovation in AI and other technologies through initiatives like the Defense Innovation Unit (DIU), an outpost of the DoD in US innovation hubs including Silicon Valley, Austin, and Boston. However, despite these ambitions, the DoD has at times been faced with hurdles establishing public-private partnerships, as companies were hesitant to collaborate with the military for economic (Olney, 2019) or moral reasons (Kuzma & Wester, 2019). For example, following a wave of protest from employees, Google decided not to renew contract with the Pentagon on Project Maven, an initiative that uses AI to support the processing, exploitation, and dissemination of video and imagery intelligence collected by drones (Wakabayashi & Shane, 2018).

3.1.2 Areas of Military AI Application

Nevertheless, and despite some setbacks, the Pentagon is already developing, testing, and even fielding the application of AI in diverse areas, including ISR, logistics, cyber, information operations, (semi)autonomous vehicles, and command and control. Currently, the US DoD is working on over 600 active AI projects (Sayler, 2020). For example, the Army's Logistics Support Activity collaborates with IBM on the optimization of the maintenance schedules for the Stryker fleet. In the area of information warfare, DARPA initiated the so-called "Media Forensics project" to counter fake news and to improve the analysis of the authenticity of visual data. Moreover, US military services seek to integrate AI into semiautonomous and autonomous vehicles. One example is the Air Force's Loyal Wingman program, "which pairs an older-generation, uninhabited fighter jet (F-16) with an inhabited F-35 or F-22" (Sayler, 2019, p. 13).

3.1.3 Risks of Military AI Applications

While the US is actively pursuing a range of AI applications, it has also indicated concerns about related ethical, legal, and safety risks. For example, the United States was the first country to develop a national policy on autonomous weapons systems. DoD Directive 3000.09 from 2012 (updated in 2017) restricts the use of autonomous weapons to the application of non-lethal, non-kinetic force (U.S. Department of Defense, 2017). However, given past statements by the US delegation at the CCW in Geneva, it is very unlikely that the US would support an international ban on LAWS. In this forum, the US delegation has repeatedly argued against a negotiation mandate and even opposed the inclusion of the notion of human control in the summary reports (Human Rights Watch, 2020).

Beyond the narrow domain of autonomous weapons systems, in its AI strategy the DoD set itself the goal of distinguishing itself as a leader in military ethics and AI safety. In this vein, in 2020 the DoD had already adopted a set of five ethics principles that are supposed to guide the US's development and deployment of military AI. To give one example, the first of these principles specifies that "DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities" (U.S. Department of Defense, 2020). However, how these principles will eventually shape AI development and deployment—especially as key US competitors move toward realizing their AI ambitions—remains to be seen. Lastly, military agencies working on AI such as DARPA have also launched programs that aim at assuring the safety and correct functioning of AI systems (Morgan et al., 2020).

3.1.4 Summary

The previous overview has shown that the US military's pursuit of AI is primarily driven by a perceived threat to its longstanding military-technological edge over competitors such as China and Russia. In order to take advantage of AI for military applications, the US is seeking to combine military R&D efforts by organizations like DARPA with public-private partnerships. Current investments and activities across multiple areas of application demonstrate the resolve of the US to make AI a focal military capability, even though it has only deployed AI in a narrow range of cases so far. The DoD has also demonstrated its concern about AI safety and legal and ethical risks and has launched several initiatives at the national level to address some of these challenges.

3.2 China

China is regarded as an emerging contender of the US in AI. In 2017, the Chinese State Council released its AI strategy—the Next Generation Artificial Intelligence Plan—setting itself the eye-catching objective of becoming the global leader in AI by 2030 (Webster et al., 2017). Currently, China leads all countries in having large datasets available required for training machine-learning algorithms, but is still behind the US regarding hardware, talent, and the quality of research publications and patents (Ding, 2018; Breitinger et al., 2020). The Chinese AI strategy is primarily concerned with applications for the civilian sector, including manufacturing, smart cities, and education, but also includes a section on defense suggesting that AI be rapidly introduced as part of its national defense innovation (Webster et al., 2017).

Chinese military AI efforts are driven by technological opportunity and strategic considerations. The Chinese People’s Liberation Army (PLA) recognizes the potential opportunities inherent in the application of AI in the military realm. However, it is also concerned about falling behind the US military, which it regards as a powerful adversary (Kania, 2020). Xi Jinping has set the target for the PLA to become a world-class military by mid-century and to finally close the capability gap with other military powers and especially the US. Since the end of the Cold War, China has invested significantly in military modernization and the development of asymmetric capabilities targeting vulnerabilities of the US military. For the PLA, the introduction of AI could herald a new era of *intelligentized* warfare and provide China with the historical opportunity not only to catch up with but also to leapfrog and surpass the US (Kania, 2018).

3.2.1 Sources of Military AI

In 2020 alone, China is estimated to have spent \$70 billion on AI, but the specific amount of defense AI R&D is unknown, as no official figures exist (Acharya & Arnold, 2019). However, China, more than any other country, blurs the line between civilian and military resources as it uses an approach to military modernization called civil-military fusion. Also in China, private sector actors lead in the development of AI, and Beijing sees these commercial developments as also significant for the military. Yet, also in China, and despite the seemingly closer link between government and the private sector, collaboration is not without friction and does not guarantee a seamless translation of civilian into military resources (Laskai, 2018; Saylor, 2020).

3.2.2 Military AI Applications

The PLA has signaled its intention to apply AI to upgrade existing weapons systems but also to develop entirely new capabilities. Chinese leaders view AI as an important element of its military modernization, but the military leadership has yet to clarify its plans and priorities (Kania, 2020). However, the PLA is already developing military AI applications across various areas, including autonomous vehicles, command decision making, ISR, cyber, and logistics. For example, the PLA is developing swarming Unmanned Aerial Vehicles (UAV) for ISR, communications and strike missions. It is also developing autonomous ground vehicles including the Sharp Claw I and II, which reportedly can “autonomously conduct reconnaissance, identify and track, and engage targets” (Morgan et al., 2020, p. 64). The Chinese Navy is supposedly also developing autonomous submarines (Kania, 2020).

3.2.3 Risks of Military AI Applications

While China is pursuing AI for military applications, Chinese leaders have also voiced concerns about the risks of an AI arms race with the US on multiple occasions (Allen, 2019). In May 2019, a group of state and non-state actors released the “Beijing AI principles” that call for cooperation in AI governance and warn against a “malicious AI race” (BAAI, 2019). At the CCW, China called for a ban on autonomous weapons systems in 2018. However, as was later clarified, China’s position was limited to a ban on the use of LAWS and did not include their development and production (Human Rights Watch, 2020). Nonetheless, China has consented to 11 guiding, non-binding principles on LAWS that were proposed by Germany and France in 2019. The principles affirm for example that a human must always be responsible for the decision to use LAWS and that international humanitarian law applies (France Diplomacy, 2019).

3.2.4 Summary

China views AI as an opportunity to move its military modernization efforts forward and to catch up with or possibly even leapfrog the US. In order to fulfill these objectives, China uses a strategy of civil-military fusion, exploiting the synergies of AI developments in the civilian and military realm. While China has not officially released a military AI strategy, the PLA is already experimenting with AI applications in various areas and has demonstrated its resolve to use it broadly in the future. However, Chinese leaders have also shown concern about potentially dangerous dynamics resulting from an AI arms race with the US and have recently called for a ban on autonomous weapons. However, it is unclear whether these concerns will actually constrain the PLA’s further pursuit of military AI.

3.3 *France*

The French President Emmanuel Macron presented the French AI strategy, which is backed by support of €1.5 billion until 2022, to the public in 2018 (Gouvernement, 2018). The overarching goal of the AI strategy is to establish France as a competitive AI power in Europe but also globally. While an increasing number of European countries have AI strategies, France stands out by also considering the military realm as vital to its AI efforts (Franke, 2019). The Villani Report, a policy report that served as the basis for France's AI strategy, had already called for a focus on the defense and security sector (Villani, 2019). Although France does not face an immediate external security threat, the strategy frames the future use of AI in the military domain as necessary to secure the country's competitiveness and security relative to other states. In addition to its national AI strategy, France, like the US, also released a military AI strategy in 2019. The authors of the military AI strategy classify France as part of "the second circle" in AI, trailing the two "superpowers", the US and China (Ministère des Armées, 2019, p. 7).

3.3.1 Sources of Military AI

In early 2018, the French Ministry of Defense (MoD) announced it was investing €100 million per year between 2019 and 2025 in AI technologies, involving all of its weapons programs (Tran, 2018). Later in 2018, France established the Defense Innovation Agency (DIA), which has been linked to the American DARPA, to foster and coordinate the application of new technology in the military realm. A dedicated Defense Artificial Intelligence Coordination Unit (CCIAD), responsible for coordinating the ministry's activity in the area of AI, has been attached to the DIA (Ministère des Armées, 2019). However, the Villani report had already also emphasized the need to exploit synergies between AI developments in the military and civilian realm (Villani, 2019). The recently established "Innovation Défense Lab" by the French MoD aims, among other objectives, to speed up the adoption of technology coming from the private sector (Ministère des Armées, 2020).

3.3.2 Areas of Military AI Applications

While the French military is still in the initial stages of AI development and deployment, France is already part of the development of different "big-ticket" systems that will involve AI technologies. The most developed project, so far, is Dassault's unmanned combat aerial vehicle (UCAV) nEUROn that, apart from France, also involves Greece, Italy, Spain, Sweden, and Switzerland. A project that is still in the early stages is the Future Combat Air System (FCAS), a joint initiative by France, Germany, and Spain, with Airbus and Dassault leading the development. Eventually, FCAS should enable "teaming between a manned fighter

and swarms of autonomous drones” (Franke, 2019, p. 16). Another project that is currently under development is dubbed Artemis and has the objective of providing the French Defense Procurement Agency with a sovereign infrastructure for the storage of big data and data management (Franke, 2019).

3.3.3 Risks of Military AI Applications

On several occasions, French decision-makers have voiced concern about the development and deployment of LAWS. French defense minister Florence Parly (2019) has stated that France had no interest in developing “killer robots” but has no national policy on LAWS in place yet. Together with Germany, France has promoted the adoption of 11 guiding principles on LAWS, the most significant outcome of the UN deliberations to date (France Diplomacy, 2019). Beyond the area of LAWS, it is noteworthy that in 2019 the French MoD set up a panel on the ethical implications of military AI applications (Parly, 2019). However, so far little is known about the work of the board, and its influence on France’s decisions concerning the use of AI in the military realm remains to be seen.

3.3.4 Summary

France is certainly one of the European countries with the most advanced thinking on military AI applications and their potential implications. The French MoD is promoting experimentation with AI technologies across different domains and has underscored this plan with substantial financial resources. Yet, French leaders are keenly aware of the leading role of the private sector in AI R&D and the necessity to seek closer collaboration. Currently, France is already involved in the development with traditional and some non-traditional defense contractors of several major weapon systems and data management systems that feature AI components. However, France also has actively engaged in discussions of the risks of military AI applications at the UN and national level and has launched first initiatives to address legal and ethical challenges.

3.4 *Israel*

Israel is already regarded as one of the world’s leading high-tech countries and, similarly to France, is also striving to position itself among the top five countries in AI. Currently, the government is working on the implementation of a five-year national AI program with a budget of \$1.55 billion (Orbach, 2020). Notably, the National Security Council and the Directorate of Defense R&D of the Israeli MoD have been closely involved in a working group on AI that was set up at the request of former Prime Minister Benjamin Netanyahu (Berkovitz, 2019).

Israel presents a special case because its army, the Israel Defense Forces (IDF), is well-known for always being on the cutting edge of high-tech and is regarded as the cradle of Israeli tech talent and the startup sector. Indeed, the IDF is often given credit for the global success of the Israeli high-tech ecosystem. The IDF's proclivity to develop and adopt novel technologies is closely linked to Israel's unique security situation in the Middle East and its frequent involvement in border disputes and wars (Swed & Butler, 2015). Due to Israel's security situation and the IDF's penchant for high-technology, it can be expected that Israel's military use of AI will increase rapidly in the near future.

3.4.1 Sources of Military AI

The IDF has already set up the so-called "Sigma branch" within its C4i technical unit, with the "purpose to develop, research, and implement the latest in artificial intelligence and advanced software research in order to keep the IDF up to date" (Israel Defense Forces, 2017). However, because of the close connection between the military and civilian sphere due to, for example, men and women completing compulsory military service, the IDF regularly collaborates with commercial companies and researchers on technology projects. Thus, in contrast to other countries, where militaries are just beginning to foster closer collaboration with the private sector in the development of military AI applications, the IDF can draw on years of experience with public-private high-tech partnerships.

3.4.2 Military AI Applications

Against this background, it comes as no surprise that Israel is already using and developing military AI applications in various areas, including but not limited to autonomous vehicles, ISR and targeting. At present, Israel already has a range of weapons with varying degrees of autonomy in the arsenal. One interesting example is the Harop (or Harpy 2) loitering munition—advertised by its developer Israel Aerospace Industries as an "all-weather autonomous weapon"—an UAV that can stay in the air for a considerable amount of time (about 6 hours) before engaging ground targets with an explosive warhead (Israel Aerospace Industries, 2020). Moreover, the IDF has been deploying unmanned vehicles for years, including the Guardium unmanned ground vehicle, which patrols the border with the Palestinian-governed Gaza Strip (Shamah, 2014). AI projects that the IDF is currently working on include the Fire Weaver sensor-to-shooter system, developed by the defense contractor Rafael. Fire Weaver relies on advanced computer vision technology and AI algorithms to facilitate precision targeting for commanders and soldiers. In a collaborative effort with startups and researchers, the IDF is also working on the Stargate and Starlink projects, which seek to leverage AI to analyze aerial and non-aerial images, respectively (Cohen, 2019).

3.4.3 Risks of Military AI Applications

Despite its broad military AI efforts, Israel does not so far have a national policy addressing the risks of military AI applications. Moreover, although Israel participates in the CCW discussion on LAWS, it has argued against a treaty or a ban on LAWS and even asked other states to “to keep an open mind regarding the positive capabilities of future lethal autonomous weapons systems” (Human Rights Watch, 2020). However, it has also agreed to the 11 guiding principles on LAWS as discussed above (France Diplomacy, 2019).

3.4.4 Summary

Due to its unique security situation, Israel has a highly trained army with extensive experience in developing high-tech programs and adopting novel technologies. Moreover, the IDF has close ties to the civilian technology ecosystem and traditional defense contractors in Israel that are among the most advanced in the world. This unique state of affairs could favor Israel in developing military AI applications and bringing them to the battlefield quickly. Israel already deploys certain AI technologies in the military realm and is expected to expand their use in the future. At the moment there are no indications that Israel will constrain its development and deployment of military AI capabilities on a national level or that it will support an international legal instrument on LAWS.

4 Comparison and Discussion

At present, only a speculative and preliminary assessment and comparison of states’ emerging military AI capabilities is possible. However, some conclusions can be drawn based on the preceding analyses. First of all, despite AI’s still emerging character and many remaining uncertainties regarding its potential, states are already showing a significant interest in applying AI for military purposes. This interest or even resolve to do so is reflected in strategic documents, including the US and French military AI strategies, the amounts spent on AI for defense, and the development of AI for military applications across several areas.

The four countries that have been assessed, independently of their size and geographic location, have in common that they view synergies between the commercial and military sector as critical to achieving their AI objectives. While some countries like the US and France may face more hurdles than China and Israel in establishing partnerships with private companies, they all face challenges in systems integration and developing AI applications specifically for military purposes.

In a few cases, systems featuring AI are already being deployed by militaries. In most cases, the development is still underway. Thus, at this point it can primarily be

concluded that the countries reviewed in this chapter intend to apply AI across a wide range of areas including logistics and training, cyber and information operations, ISR, and (semi)autonomous vehicles, as well as command and control. Whether states will be able to transform AI into truly transformative capabilities on the battlefield remains to be seen. However, while the four countries all face a unique strategic environment and the drivers for the deployment of military AI differ, they all share the perception that AI could potentially bring large military gains and that as a consequence they cannot risk falling behind.

The countries assessed in this chapter differ significantly with regard to how they approach the manifold risks of AI applications in the military. AI safety and ethics feature prominently in the US military AI strategy and France also recognizes these challenges and has started to address them. However, it is premature to say whether and how efforts like the US's AI defense principles or the French ethics panel will limit the development or use of military AI. While Israel has not yet visibly addressed the risks arising from military AI applications, China's position is more ambiguous. While Chinese leaders have demonstrated their concern regarding a potential AI arms race and the deployment of LAWS, it is difficult to decipher the country's intentions when it comes to regulation.

This chapter has made an initial attempt at reviewing and comparing the emerging military AI capabilities of four different countries. Further developments in this area should be closely monitored due to the great promise but also perils of military AI. However, the focus of the chapter has been on a small number of states that already have relatively powerful civilian AI industries and military capabilities. It is not yet clear how the technological lead the commercial AI sector has over its military counterpart will influence proliferation (Horowitz, 2018). Due to its commercial availability, less technologically and militarily advanced states could also develop military AI capabilities and bolster their relative position. AI applications in the military could also strengthen smaller states in particular because they have a smaller armies. At this point, it is too early to answer these questions, but they remain important topics for future research.

References

- Acharya, A., & Arnold, Z. (2019). *Chinese public AI R&D spending: Provisional findings*. Center for security and emerging technology, Issue Brief. Accessed May 11, 2020, from <https://cset.georgetown.edu/wp-content/uploads/Chinese-Public-AI-RD-Spending-Provisional-Findings-1.pdf>
- Alic, J. A., Branscomb, L., Brooks, H., Carter, A. B., & Epstein, G. L. (1992). *Beyond spinoff: Military and commercial Technologies in a Changing World*. Harvard Business Review Press.
- Allen, G. C. (2019). *Understanding China's AI strategy: Clues to Chinese strategic thinking on artificial intelligence and national security*. Center for a New American Security. Accessed May 11, 2020, from <https://s3.amazonaws.com/files.cnas.org/documents/CNAS-Understanding-Chinas-AI-Strategy-Gregory-C.-Allen-FINAL-2.15.19.pdf?mtime=20190215104041>

- Allen, G. C. (2017). *Putin and Musk are right: Whoever masters AI will run the world*. CNN. Accessed Mar 25, 2022, from <https://edition.cnn.com/2017/09/05/opinions/russia-weaponize-ai-opinion-allen/index.html>
- BAAI (2019). *Beijing AI principles*. Accessed Jun 15, 2020, from <https://www.baai.ac.cn/news/beijing-ai-principles-en.html>
- Barsade, I., & Horowitz, M. C. (2018). *Artificial intelligence beyond the superpowers*. Bulletin of the atomic scientists. Accessed May 11, 2020, from <https://thebulletin.org/2018/08/the-ai-arms-race-and-the-rest-of-the-world/>
- Berkovitz, U. (2019). *Israel's national AI plan unveiled*. Globes. Accessed Jan 20, 2021, from <http://en.globes.co.il/en/article-israels-national-ai-plan-unveiled-1001307979>
- Breitinger, J. C., Dirks, B., & Rausch, T. (2020). *World class patents in cutting-edge technologies*. Bertelsmann Foundation. Accessed Jun 15, 2020, from https://www.bertelsmann-stiftung.de/fileadmin/files/user_upload/BST_World_class_patents_2020_ENG.pdf
- Cohen, S. (2019). *Star Trek, Stargate and the Israeli army's other AI projects*. Haaretz. Accessed May 11, 2020, from <https://www.haaretz.com/israel-news/.premium-startrek-stargate-and-the-israeli-army-s-other-ai-projects-1.7908968>
- U.S. Department of Defense. (2020). *DOD adopts ethical principles for artificial intelligence*. Accessed Jun 18, 2020, from <https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/>
- U.S. Department of Defense. (2018). *Summary of the 2018 department of defense artificial intelligence strategy: Harnessing AI to advance our security and prosperity*. Accessed May 11, 2020, from <https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF>
- U.S. Department of Defense. (2017). *Department of defense directive 3000.09*. Accessed Jan 21, 2021, from <https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf>
- Ding, J. (2018). *Deciphering China's AI dream: The context, components, capabilities, and consequences of China's strategy to lead the world in AI*. Future of Humanity Institute, University of Oxford. Accessed May 11, 2020, from https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf
- Ellman, J., Samp, L., & Coll, G. (2017). *Assessing the third offset strategy*. Center for Strategic and International Studies. Accessed Mar 25, 2022, from https://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/170302_Ellman_ThirdOffsetStrategySummary_Web.pdf
- Esper, M. T. (2019, July 16). *United States senate, committee on armed services. Hearing to conduct a confirmation hearing on the expected nomination of: Honorable Mark T. Esper to be Secretary of Defense*. Accessed May 11, 2020, from https://www.armed-services.senate.gov/imo/media/doc/19-59_07-16-19.pdf
- France Diplomacy. (2019). *11 Principles on lethal autonomous weapons systems (LAWS)*. Accessed Jun 15, 2020, from <https://www.diplomatie.gouv.fr/en/french-foreign-policy/united-nations/alliance-for-multilateralism-63158/article/11-principles-on-lethal-autonomous-weapons-systems-laws>
- Franke, U. E. (2019). *Not smart enough: The poverty of European military thinking on artificial intelligence*. European Council on Foreign Relations, Policy Brief. Accessed May 11, 2020, from https://www.ecfr.eu/page/-/Ulrike_Franke_not_smart_enough_AI.pdf
- Gouvernement. (2018). *Artificial intelligence: Making France a leader*. Accessed Jun 10, 2020, from <https://www.gouvernement.fr/en/artificial-intelligence-making-france-a-leader>
- Hagel, C. (2014). *Secretary of defense speech: Reagan national defense forum keynote*. Accessed Jun 12, 2020, from <https://www.defense.gov/Newsroom/Speeches/Speech/Article/606635/>
- Hao, K. (2019). *Yes, China is probably outspending the US in AI—but not on defense*. MIT Technology Review. Accessed May 11, 2020, from <https://www.technologyreview.com/2019/12/05/65019/china-us-ai-military-spending/>
- Harper, J. (2020, 3. February). *Analysts say \$25 billion needed for AI*. National Defense. Accessed Jun 12, 2020, from [https://www.nationaldefensemagazine.org/articles/2020/2/3/analysts-say-\\$25-billion-needed-for-ai](https://www.nationaldefensemagazine.org/articles/2020/2/3/analysts-say-$25-billion-needed-for-ai)

- 5-billion-needed-annually-for-ai#:~:text=Defense%2Drelated%20AI%20spending%20was,to%20invest%2C%20the%20analysts%20said
- Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*, 1(3), 36–57. <https://doi.org/10.15781/T2639KP49>
- Horowitz, M. C., Kania, E. B., Allen, G. C., & Scharre, P. (2018). *Strategic competition in an Era of artificial intelligence*. Center for a New American Security. Accessed May 11, 2020, from <https://www.cnas.org/publications/reports/strategic-competition-in-an-era-of-artificial-intelligence>
- Human Rights Watch. (2020). *Stopping killer robots: Country positions on banning fully autonomous weapons and retaining human control*. Accessed Jan 22, 2021, from <https://hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and#>
- Israel Aerospace Industries. (2020). *Harpy: Autonomous weapon for all weather*. Accessed May 11, 2020, from <https://www.iai.co.il/p/harpy>
- Israel Defense Forces. (2017). *The IDF sees artificial intelligence as the key to modern-day survival*. Accessed May 11, 2020, from <https://www.idf.il/en/minisites/technology-and-innovation/the-idf-sees-artificial-intelligence-as-the-key-to-modern-day-survival/>
- James, A. (2013). *Emerging technologies and military capabilities*. S. Rajaratnam School of International Studies, Policy Brief. Accessed Jun 15, 2020, from https://www.rsis.edu.sg/wp-content/uploads/2014/07/PB131101_Emerging_Technologies_and_Military_Capability.pdf
- Kania, E. B. (2020). “AI weapons” in China’s military innovation. Center for Security and Emerging Technology. Accessed Jun 18, 2020, from https://www.brookings.edu/wp-content/uploads/2020/04/FP_20200427_ai_weapons_kania_v2.pdf
- Kania, E. B. (2018). *China’s strategic ambiguity and shifting approach to lethal autonomous weapons systems*. Lawfare. Accessed Jun 10, 2020, from <https://www.lawfareblog.com/chinas-strategic-ambiguity-and-shifting-approach-lethal-autonomous-weapons-systems>
- Kuzma, R., & Wester, T. (2019). *Economics sure, but don’t forget ethics with artificial intelligence*. Strategy Bridge. Accessed May 11, 2020, from <https://thestrategybridge.org/the-bridge/2019/3/5/economics-sure-but-dont-forget-ethics-with-artificial-intelligence>
- Laskai, L. (2018). *Civil-military fusion: The missing link between China’s technological and military rise*. Council on Foreign Relations. Accessed Jun 10, 2020, from <https://www.cfr.org/blog/civil-military-fusion-missing-link-between-chinas-technological-and-military-rise>
- Leung, J., & Fischer, S.-C. (2018). *JAIC: Pentagon debuts artificial intelligence hub*. Bulletin of the Atomic Scientists. Accessed May 11, 2020, from <https://thebulletin.org/2018/08/jaic-pentagon-debuts-artificial-intelligence-hub/>
- Ministère des Armées. (2020). *Innovation défense lab*. Accessed Jan 20, 2021, from <https://www.defense.gouv.fr/aid/actualites/innovation-defense-lab>
- Ministère des Armées. (2019). *Artificial intelligence in support of defence*. Report of the AI Task Force. Accessed May 11, 2020, from <https://www.defense.gouv.fr/sites/default/files/aid/Report%20of%20the%20AI%20Task%20Force%20September%202019.pdf>
- Morgan, F. E., Boudreaux, B., Lohn, A. J., Ashby, M., Curriden, C., Klima, K., & Grossman, D. (2020). *Military applications of artificial intelligence: Ethical concerns in an uncertain world*. RAND Corporation. Accessed Jun 20, 2020, from https://www.rand.org/pubs/research_reports/RR3139-1.html
- Olney, R. (2019). *The Rift between Silicon Valley and the pentagon is economic, not moral*. War on the Rocks. Accessed May 11, 2020, from <https://warontherocks.com/2019/01/the-rift-between-silicon-valley-and-the-pentagon-is-economic-not-moral/>
- Orbach, M. (2020). *Israel launches national AI program, but lack of budget threatens its implementation*. CTECH. Accessed Jan 19, 2021, from <https://www.calcalistech.com/ctech/articles/0,7340,L-3883355,00.html>
- Parly, F. (2019). *France will not develop “killer robots”*—Speech by the Minister of the Armed Forces on AI. Permanent representation of France to the Conference on Disarmament. Accessed

- Jun 20, 2020, from <https://cd-geneve.delegfrance.org/France-will-not-develop-killer-robots-Speech-by-the-Minister-of-the-Armed>
- Saslow, K. (2020). *Understanding US federal AI policy: recommitting to a transatlantic coalition on AI. Memo.* Stiftung Neue Verantwortung. Accessed Jan 22, 2021, from <https://www.stiftung-nv.de/en/publication/understanding-us-federal-ai-policy-recommitting-transatlantic-coalition-ai>
- Sayler, K. N. (2020). *Emerging military technologies: Background and emerging issues for congress.* Accessed Jan 23, 2021, from <https://fas.org/sgp/crs/natsec/R464588.pdf>
- Sayler, K. N. (2019). *Artificial intelligence and national security.* Congressional Research Service. Accessed May 11, 2020, from <https://fas.org/sgp/crs/natsec/R45178.pdf>
- Scharre, P. (2020). *The militarization of artificial intelligence.* Texas national security review. Accessed Jun 19, 2020, from <https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security/#essay4>
- Shamah, D. (2014). *As google dreams of driverless cars, IDF deploys them.* The times of Israel. Accessed May 11, 2020, from <https://www.timesofisrael.com/as-google-dreams-of-driverless-cars-idf-deploys-them/>
- Swed, O., & Butler, J. S. (2015). Military Capital in the Israeli hi-tech Industry. *Armed Forces & Society*, 41(1), 123–141. <https://www.jstor.org/stable/48609201>
- The White House. (2017). *National security strategy of the United States of America.* Accessed May 11, 2020, from <https://www.whitehouse.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905.pdf>
- Tran, P. (2018). *France to increase investment in AI for future weapon systems.* Defense News. Accessed May 11, 2020, from <https://www.defensenews.com/intel-geoint/2018/03/16/france-to-increase-investment-in-ai-for-future-weapon-systems/>
- Villani, C. (2019). *For a meaningful artificial intelligence: Towards a French and European strategy.* Accessed May 11, 2020, from https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf
- Wakabayashi, D., & Shane, S. (2018). *Google will not renew pentagon contract that upset employees.* New York Times. Accessed Jun 17, 2020, from <https://www.nytimes.com/2018/06/01/technology/google-pentagon-project-maven.html>
- Webster, G., Creemers, R., Triolo, P., & Kania, E. (2017). *Full translation: China's 'New generation artificial intelligence development plan'.* Accessed Jun 15, 2020, from <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/>

Artificial Intelligence as an Arms Control Tool: Opportunities and Challenges



Niklas Schörnig

● آقای هوش مصنوعی ●

📰 رسانه هوش مصنوعی دانشگاه تهران 📰

@MrArtificialintelligence

Abstract In arms control, verification is the essential mechanism that ensures compliance with a treaty or regulation. However, verification is not always an easy task, especially when the contracting parties are suspicious of each other. This text shows in a systematic way how AI can promote verification in the future and presents several projects currently in different stages of development. Starting with how AI-aided translation and analysis of text can support the work of inspectors, the chapter continues to look at the analysis of graphical data, other sensory data, and the possibilities to include multimodal data into the analysis. Many of the projects presented have already passed the proof-of-concept phase and could be deployed in the next few years. However, the text emphasizes the need to use AI only in a team with human inspectors and it calls for more collaboration between AI experts and arms control experts to fully exploit the potential that AI offers for verification.

1 Introduction

Technology and intelligence have always played an important role in and advanced the field of modern arms control (Jasani & Barnaby, 1984), e.g. satellites, surveillance aircraft or improved sensor equipment for detecting traces of radiation or chemical agents. More recent technologies, which have also come to be known under the sometimes misleading heading “emerging technologies”—especially drones or other uncrewed vehicles—are also in the process of enhancing verification. Uncrewed surface vehicles to “help IAEA inspectors verify the presence of nuclear material stored underwater” (Silva & Klingenboeck, 2019) or drones for environmental monitoring are straightforward applications of emerging technologies for arms control purposes and first trials have demonstrated the usefulness of the systems. The most promising, yet significantly more difficult application will be to make artificial intelligence (AI) in general and machine learning (ML) in particular

N. Schörnig (✉)
Peace Research Institute Frankfurt, Germany
e-mail: schoernig@hsfk.de

useful in arms control. The fact that many projects have presented proof of general concepts over the last few years (see examples below) and that some international organizations have been pursuing the use of machine learning to assist human inspectors and analysts goes hand in hand with the overall progress which has been made in the realm of machine learning. However, the idea of using algorithms for arms control is not new. A SIPRI volume on Arms and Artificial Intelligence (Din, 1987) dedicates a whole section (Part IV) to “Applications [of AI] in arms control analysis.” While the AI described essentially amounted to deterministic expert systems, as machine learning based on neural networks as we know it today had just been developed and was too computationally intensive for the available computing power at that time, the book still offers interesting insights. One of them is that arms control can benefit from the implementation of AI. This is as true today as it was in 1987. The key difference, however, is that significant advantages in ML, paired with the ever-increasing computing power of modern microchips, has led to a situation where the application of AI in arms control has left the field of mere theory. We are at the beginning of an AI revolution and arms control is one field where AI experts can run into technical challenges while promoting a worthwhile cause at the same time. The aim of this chapter is to build a bridge between AI experts and arms control specialists.

The text starts by reviewing the fundamental ideas in arms control and verification. Since most readers from the arms control community will be familiar with these ideas, the section is primarily intended for AI experts wishing to understand the theoretical and conceptual background of arms control. The next section presents several ongoing projects. Some of the examples of how AI is used for arms control purposes can be found in the various chapters of this volume, often in greater detail. The aim of this chapter is to more systematically classify the examples of AI-enhanced arms control based on the data used in the case examples. The categorization used by Gastelum et al. (2018) is adopted, but another category, the analysis of sensor data other than text or images, is added. After presenting successful cases, section four looks at the problems and pitfalls that are sometimes described by analysts or sometimes ignored despite their being in plain sight.

The final section returns to the fundamental question regarding arms control: Where will AI help to foster arms control, where do the limits lie, and what is politically necessary to make more use of AI in arms control.

2 Theory of Arms Control, Disarmament and Non-proliferation

2.1 *Arms Control, Disarmament and Non-proliferation: Basic Concepts*

When it comes to arms control, several key concepts must first be clarified. First, there is a significant difference between arms control and disarmament. Disarmament always entails a notion of actually reducing armament (disarmament as an ongoing process) or even achieving a state of zero armament, meaning that full disarmament (disarmament as a final status) and the banning of a certain weapon (Carter, 1989, p. 4). Arms control, by contrast, can be understood in a broader and less strict way, including, for example, agreeing on certain ceilings for a particular weapon system or even controlled armament, that is to say, a mutually agreed upon limitation of build-up rates. Proponents of disarmament see a direct connection between the number of weapon systems and the potential for conflict, arguing that less is always better (Müller & Schörnig, 2006, pp. 124–126; see Müller, 2017, p. 11 for an English version). From this perspective, disarmament is an end in itself. The most important aim of arms control is avoiding (nuclear) war in a conflictual international environment (Schelling & Halperin, 1961, p. 1). But instead of complete disarmament, arms controllers aim for stability based on a military balance as the best protection against war (Schelling & Halperin, 1961, p. 1; Carter, 1989, p. 4). Agreeing on a long-term balance of power, for example, by agreeing on mutual rates of weapons build-up, would be seen as better than unilateral but destabilizing disarmament, at least in a competitive environment. But arms control also aims to limit the costs of armament (e.g., through a freeze without an actual reduction) and the limitation of destruction in war (Schelling & Halperin, 1961, p. 1). Some authors also stress the importance of confidence and security building and other transparency-creating measures in the context of arms control, as these measures usually help to stabilize relations between rational adversaries (Altmann, 2019).

Non-proliferation, to add the third relevant concept, focuses on a distinction between haves and have-nots, where those states already in possession of a particular weapon system refuse to give it to states not yet in possession, thus limiting its proliferation. The concept historically refers to the nuclear realm (Goldblat, 1982, p. 45). A related, yet less strict concept is that of export controls, where those in possession of a certain weapon system or militarily usable technology choose not to sell to every potential buyer, or only export subject to certain restrictions (for the case of drones, for example, see (Schörnig, 2017)) such as the recipient state not being allowed to pass on or sell the system to a third party or limiting the capabilities of the system.

In all cases, however, one of the most pressing questions is whether states, or in the realm of non-

proliferation and export control, both state and non-state actors adhere to an agreement. Especially when the security of a state is at stake, trust in and the

reliability of an agreement is essential and assurance that all members will comply with an arms control agreement and no one will cheat is the “state’s most serious concern” (Karkoszka, 1977, p. 8). As a result, verification is needed to demonstrate compliant behavior.

2.2 *Verification*

Verification is understood as the essential mechanism that ensures compliance with a treaty. The often-quoted tagline is an old Russian proverb, frequently used by Ronald Reagan: “Trust, but verify!” And “there is a consensus that verification needs to be built into an [international arms control] agreement” (Keir & Persbo, 2020, p. 16)—with the Biological Weapons Convention being the exception (see the text by Filippa Lentzos in this volume). In practice, verification can, for example, determine whether a weapon system’s treaty-defined ceiling, for example the number of missiles in an installation or of tanks in a specific region, is in compliance with the agreement. Or verification can check whether a specific factory is showing traces of the production of forbidden chemicals or the enrichment of uranium.

However, in a way the term “verification” is misleading as it suggests that compliant behavior could in fact be “verified” or “confirmed”, ruling out the possibility of violations of an agreement, or “cheating.” However, this would overburden verification as no arms control treaty is safe against skillful and resource-intensive cheating. The aims of verification or so-called safeguard measures are therefore less ambitious. They include, among others, first, to “ensure that violations are not likely to remain covered,” (Gayler, 1986, p. 5), or concealed and deter the potential violator as a result (Goldblat, 1982, p. 90). Second, to raise the costs of treaty violations significantly, and thus reducing incentives to cheat; third, to create such a dense net of verification measures as an early warning mechanism that actual violations of a treaty are detected before their impact on security becomes severe, thus leaving enough options for reacting (Gayler, 1986, p. 4), and fourth, although there may be residual uncertainties, to increase trust and confidence between adversaries over time when no violations are detected (Gayler, 1986, p. 4; Goldblat, 1982, p. 90). Verification comes in many forms and has highly technical aspects. It starts with unilateral measures based on so-called national technical means (NTM), such as satellite observation of relevant installations or intelligence gathering and analysis (Gayler, 1986, p. 4). Thus, even if the treaty does not provide a specifically designed verification regime, states can monitor their peers’ behavior and impose consequences when necessary. In some cases, treaties explicitly mention national technical means as a verification mechanism and prohibit actions which interfere with verification, such as concealing missile sites. The 1972 Strategic Arms Limitations Talks/Treaty (SALT) between the United States and the Soviet Union are a case in point here (Goldblat, 1982, pp. 90–91).

More formalized forms of verification include, among others, the exchange of protocols, numbers and otherwise relevant material, the monitoring of military drills

and maneuvers, routine inspections of agreed locations or even surprise or “challenge” inspections with only limited warning time.¹ In some cases, the verification is conducted by experts from the parties to a treaty, in others information is gathered by independent inspectors on-site or via technical means run by an international agency such as the International Atomic Energy Agency (IAEA) or the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO).

It is obvious that verification walks a tightrope. If the verification procedure is too lax, it is considered to have limited value, if any. If it is too intrusive, the inspecting country may gain additional sensitive information not related to the original intent of the verification mechanism. It is no wonder that the accusation that verification measures could also be used for espionage—or even that they are a legalized form of spying—has accompanied the debate for decades and can still poison treaties that are otherwise working well. The unsubstantiated accusation in the debate about the US withdrawal from the Open Skies Treaty in 2019/2020 is a case in point here.²

As mentioned above, advocates stress that successful verification of certain facts, for example the verification of a certain number of a specific weapon system at a specific place as reported, increases confidence in the material received, especially after several rounds of iteration. Practice has shown that in addition to this there is yet another layer: the personal interactions between inspectors from different, possibly even hostile states may lead to personal relations and trust, and thus provide trustworthy contact persons in a crisis with whom tensions can be de-escalated on both sides.³

And finally, verification also helps states wrongly accused of cheating prove their compliance with treaties (Lück, 2019, p. 16), at least to a certain extent, as very paranoid treaty parties (or factions within a treaty party) can always assume the existence of a secret facility not detected by verification measures, or imagine some other form of cheating.

2.3 AI-Enhanced Verification from a Theoretical Perspective

When it comes to the practical implementation of verification measures, the process of verification usually cycles through three phases: first, the gathering of information; second, reviewing the information; and third, determining compliance based on the information gathered and reviewed (Keir & Persbo, 2020, p. 17). It is obvious that AI can be of huge help in this process, either by adding deeper layers of analysis to existing data or by tapping into new sources of information. In the case of

¹Excellent overviews by Karkoszka (1977, pp. 15–37) or Jasani and Barnaby (1984) are still pertinent today.

²<https://www.cotton.senate.gov/news/press-releases/cotton-cruz-introduce-resolution-calling-for-withdrawal-of-united-states-from-open-skies-treaty>, retrieved 18 February, 2022.

³Personal conversation with people involved in actual inspection measures (November 2018).

treaty-based verification measures, the role of AI may be limited, as the pool of sources is often defined in the treaty. This might not be problematic in other forms of verification, either when verification measures are enforced upon an entity (via a UN Council resolution), when a state monitors national companies' and corporations' behavior (as in the case of export controls) or when civilian entities, such as NGOs or scientists, offer a form of societal verification⁴ based on the use of publicly available sources to monitor compliance or non-compliance of state actors.

But even within existing verification regimes, it is obvious that arms control in general and verification in particular involve tasks where a great deal of material has to be assessed while paying attention to many details at the same time. In many regimes, verification measures still go back to the time when the regime was created, such as in the case of the Chemical Weapon Convention in the early 1990s, and there is a general need to update the instruments: "It is essential to consider new technologies in order to guarantee efficient verification of non-proliferation and disarmament treaty compliance for the future" (Schulze et al., 2020, p. 188). However, formal mechanisms for adapting a treaty or its protocols are already included in several verification regimes.

While simpler deterministic AI algorithms or expert systems can be helpful in assisting arms-control efforts by helping experts or inspectors to not lose oversight in a complex situation, machine learning can—as in other domains as well—once again significantly enhance the effectiveness and reliability of analyses. It is obvious that AI cannot be a help when it comes to fostering interpersonal relations. But it can be a significant help in the technical analysis of data. Only time will tell whether it is "merely" an improvement or the much-vaunted gamechanger. But it is certain that it will be a help. "Broadly, machine learning is the art of teaching a machine to make predictions based on past observations" (Gastelum & Shead, 2018, p. 42). A major aspect of verification efforts involves predicting the chances of a significant violation by another actor, finding mismatches or contradictions in the data presented and the material obtained. This is what AI is particularly good at: "Machine learning and AI can identify which data is most important, identify patterns and anomalies and predict response to certain actions" (Williams, 2020, p. 9). How this theoretical reflection fits into existing verification practices, however, will be debated at a later point in this chapter.

3 AI for Enhancing Arms Control and Verification

3.1 *Translation and Analysis of Text*

The first and most obvious hurdle when inspecting a huge number of countries is that the number of experts with adequate knowledge of the language of each country is

⁴See for example Gastelum (2020) for the concept.

limited. For example, while many Western arms control experts tried to acquire at least some proficiency in Russian during and immediately after the Cold War, arms control experts who can communicate fluently about very technical matters in Chinese are probably rare, not to mention those who speak Farsi or Korean. AI-supported translation services such as Google Translate or DeepL are still unable to achieve what specifically trained human translators are able to when it comes to complex or nuanced text. Nevertheless, these translation services have improved significantly in recent years. In many cases many translations have reached a satisfactory level of accuracy, depending on what is needed. Making it possible for technical arms control experts to understand open-source material such as newspaper articles, government statements, social media content or other material in a language they are not familiar with is a valuable asset which should not be underestimated. Institutions like DARPA, the US Defense Advanced Research Projects Agency, have identified AI-supported language translation as a very important field in which to invest. Its Broad Operational Language Translation (BOLT) program, started as early as 2011, aimed at “enabling communication with non-English-speaking populations and identifying important information in foreign-language sources.”⁵ While DARPA focused on the ability of US soldiers to interact with local populations, the idea that inspectors should have the ability to communicate without having to rely on interpreters also seems to have merits. Other known DARPA programs aimed at the translation of lesser-known languages (Translate Any Language⁶) or the automated translation of foreign language text into English, include a project called “Multilingual Automatic Document Classification, Analysis and Translation” (MADCAT),⁷ originally designed to translate captured material found in enemy hideouts. It is obvious that inspectors could also benefit immensely from an app translating photographed text in operational situations.⁸ But communication is not a one-way street. In addition to using translation tools to understand source material, important players such as international organizations could use AI to translate their publications and reports into more than the usual number of languages, thus raising awareness and increasing outreach. While publications would definitely need to be brushed up by professional linguists, the workload would be significantly reduced. AI can also help to make material accessible to begin with: Extracting coherent and processable text from PDFs or scans seems to be rather hard when the

⁵ <https://www.darpa.mil/program/broad-operational-language-translation>, retrieved 18 February 2022.

⁶ <https://slator.com/technology/darpa-does-out-millions-to-academia-and-vendors-to-translate-any-language-by-2019/>, retrieved 18 February, 2022.

⁷ <https://www.darpa.mil/program/multilingual-automatic-document-classification-analysis-and-translation>, retrieved 18 February, 2022.

⁸ According to DARPA, MADCAT has “developed optical character recognition and machine translation capabilities for 11 languages: Arabic, Chinese, Dari, Farsi, Hindi, Pashto, Spanish, Russian, Thai, Urdu and Korean”. <https://www.darpa.mil/program/multilingual-automatic-document-classification-analysis-and-translation>, retrieved February 18, 2022.

text flow is broken by the layout, for example when in a two- or three-column format, and new algorithms would be helpful here (Bast & Korzen, 2017).

AI can also help to analyze text or visualize existing data in new forms, and this makes it easier for humans to grasp specific interconnections or important aspects. Versino et al. (2018, p. 10) remind us that officials “often . . . lack the analytic and visualization tools to glean meaningful, actionable data from very large datasets.” In the realm of export controls, for example, specifically trained algorithms could link information from different sources of dual-use material to a specific location. While each individual delivery would not raise an alarm, assessing them together over time and space might.

Advanced search engines use “faceted search” to narrow down unspecific inquiries leading to too many results by including additional specifications suggested by the engine to fit the user’s needs. AI can help individual users by analyzing and taking into account their search history and connecting specific interests or drawing attention to links to new aspects the user did not originally notice.⁹ “Fuzzy search” is also an important aspect, by means of which AI will help arms controllers. Fuzzy search looks at similarities, rather than exact matches (Bast & Celikik, 2010), considering results for, say, a person’s name which are similar enough to be of interest, such as displaying results like “Schörnig,” “Schornig” or “Schörning” after using the search term “Schörnig” when other aspects (location, topics, etc.) also seem to fit. On the other hand, the same fuzzy search might also exclude exact matches when other aspects seem to indicate that the result is not appropriate (e.g., in the case of a “Michael Smith”). More advanced systems might even match quite different spellings based on their similar pronunciation.¹⁰ Fuzzy search can help arms controllers to assess material where typographical errors and/or different notations occur (e.g., translating from one alphabet into another)—carried out across different databases. This might be of special interest in the area of non-proliferation, where thousands of export-related documents which are prone to smaller mistakes and deviations are generated every day, making the task of finding relevant links in the sheer volume of documents (see Steward et al., 2018) even harder, but other applications, such as the analysis of newspaper or other open-source material, also come to mind.

3.2 *Analysis of Graphical Data*

Graphical data plays an important role in arms control and verification. Checking satellite images for changes in the landscape, in the size of an installation or in the number of weapon systems visible on a base or in a certain area is one of the most

⁹For a good overview of different search algorithms, see Steward et al. (2018, p. 24f).

¹⁰<https://medium.com/data-science-in-your-pocket/phonetics-based-fuzzy-string-matching-algorithms-8399aea04718>, retrieved February 18, 2022.

common methods of unilateral verification using a national technical means. As Jacques Baute, Director of Safeguards Information Management at the IAEA explains: “The analysis of satellite images provides an effective means to assess the nature, extent, and technical connections of nuclear sites” and adds that “the prospect of contribution from machine learning remains promising” (Baute, 2018, p. 6). In addition, safeguards regarding the non-proliferation of nuclear material entail, among other things, constant video surveillance of critical locations to detect and deter materials diversion at a specific facility. Going through a large number of pictures to check for slight differences or identifying a certain weapon system on a rather blocky or blurred picture is a demanding task which requires intense concentration. When it comes to video feeds, large parts, however, are “static” and “events of interest are infrequent,” as a workshop report of the International Atomic Energy Agency describes (IAEA, 2020, p. 10). Having to watch all footage and set the noteworthy frames apart from the uninteresting is a demanding yet boring job, or a “tedious and labor-intensive task” as the IAEA describes (IAEA, 2020, p. 10). In addition, the number of available satellites and the availability of cheap, yet powerful digital cameras has increased the sheer number of sources significantly. It is obvious that AI can be a significant help here, as the civilian application of machine learning for picture analysis and object classification has made tremendous advances in recent years. As Gastelum and Shead argue: “Some of the best-known applications of neural networks are for image classification tasks, in which the output features of a network encode a fixed set of labels and the network predicts which label(s) apply to an input image” (Gastelum & Shead, 2018, p. 42). Google Vision is a prime example of an Application Programming Interface (API) for object recognition and “image understanding,” where users can upload photos to identify objects, a person’s mood or text, based on a pre-trained machine learning model or the users’ custom models.¹¹ Inspectors could, for example, use this technology to identify export-controlled items by comparing a picture of potential contraband with a database of prohibited goods, obtain understanding of the function of a certain object based on comparable photos (Steward et al., 2018, p. 27f) or receive information indicating where or when a picture was taken.¹² A very similar project was presented by Jamie Withorne (2020), where a machine learning model was trained to identify restricted dual-use goods for non-proliferation purposes. Finally, examples of the concrete use of AI-based image analysis for arms control issues in a broader sense have included identification of nuclear facilities in operation based on their cooling towers and the emission of steam plumes based on Flickr images (Gastelum & Shead, 2018), distinguishing of unproblematic copper mills from proliferation-relevant uranium mills (Sundaresan et al., 2017) or identification of smuggled small arms or other

¹¹ <https://cloud.google.com/vision>, retrieved January 19, 2021. Unfortunately, Google has since removed the option to try out the technology without registering.

¹² In one sample picture of my mother visiting Berlin in the 1960s her clothes were correctly, yet somewhat vaguely identified as “vintage” and “retro style”.

prohibited devices on cluttered pictures of x-rayed shipping containers entering or leaving port (Rogers et al., 2017).

Another important aspect, as mentioned above, is the identification of changes in pictures taken at different points in time. Has a specific installation been expanded? Are there indications of freshly moved soil? Has a new weapon system such as a combat drone been transferred to a specific facility?

The idea of using commercial satellites to support existing verification regimes has been debated for some time now.¹³ Using AI and ML to support the analysis seems obvious. Rutkowski et al. (2018), for example, developed an algorithm analyzing Synthetic Aperture Radar (SAR) satellite images collected by ESA and released via the Google Earth Engine with the explicit aim of making a case for supporting “international verification regimes by offering a remote mechanism for treaty verification” (Rutkowski et al., 2018, p. 48). However, algorithms used in such cases need to “understand” the difference between human-made changes and changes due to lightning, different weather conditions or the angle of the image. Nonetheless, many experts are confident that these problems have already been solved to a satisfactory degree.¹⁴

Finally, when it comes to video footage, many hours may be of no interest, and only a fraction of the material may be of relevance. Algorithms identifying static pictures at a higher speed than a human can when fast-forwarding (with the risk of missing relevant sections when fast-forwarding too quickly) can flag relevant sections for human inspectors to watch more closely. The IAEA “has already been pursuing learning-based algorithms to help automate detecting and tracking events of interest” (IAEA, 2020, p. 10).

3.3 *Analysis of Other Sensor Data*

In addition to the analysis of text (spoken or written), images or film, arms control experts often have to deal with data provided by sensors. Sensors indicating the presence of toxic substances or the vibration caused by either an earthquake or a nuclear-test explosion are cases in point. Depending on the data under scrutiny, interpreting the results can be simple (reading a meter) or very challenging (interpreting seismic data). In any case, AI can assist experts in the analysis and interpretation of such data.

The often-cited suggestion of collecting telemetric data from drones to check for unwanted use of autonomous behavior in weapon systems (Gubrud & Altmann, 2013; see also the text by Dahlmann in this volume) could be enhanced by

¹³<https://vcdnp.org/emerging-satellites-for-non-proliferation-and-disarmament-verification/>, retrieved February 18, 2022.

¹⁴I received similar comments assessing this from the anonymous reviewer as well as from Thomas Reinhold (personal communication, October 2021). See also Molinier et al. (2007).

AI-supported analysis to check for deviations from expected behavior not noticeable by a human analyst. As early as 2010, Russel, Vaidya and Le Bras suggested using machine learning to enhance Comprehensive Nuclear-Test-Ban Treaty (CTBT) monitoring (Russel et al., 2010); see also the text by Anna Heise in this volume). While the CTBT has not entered into force yet, it already has its own verification organization, the CTBTO. It runs a network of more than 300 sensor stations across the globe with complementary verification technologies, the International Monitoring System (IMS),¹⁵ of which 170 are dedicated to seismic monitoring.¹⁶ All data is collected in real time at the International Data Center (IDC) and checked, for example, for significant seismic activity caused by a forbidden nuclear explosion, as opposed to natural seismic activity such as an earthquake. While some pre-selection of data was already automated by the time of writing, Russell et al. (2010, p. 32) argued that “incorporating machine learning methods into the IDC framework could improve the detection and localization of low-magnitude events, provide more confidence in the final output, and reduce the load of the human analyst.” In addition, algorithms could also support inspectors. After the CTBT enters into force, the CTBTO can send an inspection team into an area where a potential test occurred, setting up seismic equipment to detect faint aftershocks (Altmann, 2020, p. 237), which can be masked by disturbances (Altmann, 2020, p. 239). Jürgen Altmann, however, goes a significant step further and argues that acoustic and seismic sensors could be used for early warning activities such as monitoring movement of military vehicles in peacekeeping operations, at cease-fire lines, or in weapon-free zones (Altmann, 2020, pp. 240–241). This idea actually goes back three decades when the idea of classifying heavy land vehicles by sound and vibration was developed and tested in early and later experiments (Hochmuth et al., 2001; Altmann et al., 2002). Obviously, modern machine learning algorithms could enhance detection significantly, helping to detect real dangers and separate them from false alarms or false negatives (Lück, 2019, p. 20f). Altmann (2020, p. 241) also proposes detecting launches of intercontinental ballistic missiles via pre-installed geophones [or: seismic sensors] around known launch sites, or for improved early warning by signaling additional confirmation for true negatives. Given the extreme loudness of a missile launch, however, it is likely that a launch would be detected by simple measurement, so that algorithmic evaluation of the signal does not seem to be so important.

¹⁵<https://www.ctbto.org/verification-regime/background/overview-of-the-verification-regime/>, retrieved 18 February, 2022.

¹⁶<https://www.ctbto.org/verification-regime/monitoring-technologies-how-they-work/seismic-monitoring/>, retrieved February 18, 2022.

3.4 *Multimodal Data and Other Uses*

All examples described above focus on one particular set or class of data. Inferences are drawn, for example, either based on the analysis of text or specific images or other very specific sensor data. However, analysis based on multimodal data—that is, combining data based on text, images and video—brings together different data but looks at it with the same aim in mind. Feldman et al. present a “large-scale multimodal retrieval system to help analysts triage and search open source science, technology, and new data for indicators of nuclear proliferation capabilities and activities” (Feldman et al., 2018, p. 68). It is certainly true that this approach promises yet another increase in performance when information from very different sources, all addressing one particular research question, is considered. Brase, McKinzie and Zucca (2020, p. 300) conclude that “large scale integrated data models will allow analysts to answer state-specific questions and better track the evolution of potential proliferation activities.” However, not only do potential problems multiply (see below) but, given that many single-source approaches are still in their infancy, reliable working analyses based on multimodal data analysis should only be expected in the medium or long term rather than the short term.

4 **Limitations and Challenges**

Many of the examples described in the last section (and other texts in this volume) have not been applied in practice, but describe options, possibilities, proofs of concepts or first-stage prototypes. As in civilian industry, which promised, for example, that self-driving cars would be on the market years ago, the devil is often in the detail. One of the major problems AI faces when applied in the realm of arms control and verification is that the events AI has to be trained for are very rare but have very serious consequences. Knowledge and expertise from autonomous driving can indeed be incorporated in situations involving unexpected and rare events with very serious consequences (IAEA, 2020, p. 11). While the automobile industry is lagging behind expectations, as full automation in a dynamic and often unstructured environment (driving) has proven to be more difficult than expected, many use cases in the arms control realm refer to a structured and clearly delineated setting. In addition, as described in the theory section, verification does not aim at perfect security, but at least for sufficient security to avoid unpleasant surprises. However, most AI algorithms would strive to detect a “first occurrence”—the first significant violation of an agreement or the first significant deviation from the norm—as far as we know. Obvious and clear-cut cases of treaty violations are rare events. This is reflected in the datasets based on which the algorithms will or would be trained. Would an AI distinguish a dog from a cat if it had only been trained with a large number of pictures of dogs and only a few of cats? In any case, experts agree that when it comes to machine learning, the “key challenge in classification relates to the

training of algorithm[s]” (Steward et al., 2018, p. 26). Training arms control algorithms would be no exception.

Some experts state that over 90% of their time is spent preparing and managing data to make it useful for their algorithms (IAEA, 2020, p. 11). While many datasets are publicly available, including labeled imaging sets, there is still the question of how much data is enough to reliably train an algorithm. In many of the test cases described above the experts used from a few hundred to a few thousand examples to train their concepts. Would that be enough to bet a country’s security on, when other experts claim that learning algorithms need “millions of samples”¹⁷ for them to perform properly? And will there be enough publicly available data in the future with which to train algorithms? Some argue, for example, that Western states are at a disadvantage due to privacy concerns compared with other, more relaxed actors. While data to sharpen algorithms seems endless at first glance, Rose Gottenmeller (2020, p. vi) reminds us that we “cannot assume that information will always be so readily available” or that it will be reliable enough to be used. The argument has a positive aspect, however: those datasets collected and labeled by experts from international organizations such as the IAEA, the CTBTO or the [Organisation for the Prohibition of Chemical Weapons \(OPCW\)](#) should enjoy a high level of trust and reliability issues should be minimal given that multinational and impartial teams are involved.

A final problem is related to the “black box” character of self-learning AI. If, for example, an AI signals non-compliance which is totally contrary to the impression of seasoned inspectors, the reliability of the AI might be questioned for technical as well as for political reasons.

States would not only be reluctant to base their arms control decisions on unknown algorithms or databases but might fear new options capable of hiding manipulation and cheating in a tremendous amount of data. Can data sets or algorithms perhaps be manipulated in such a way that an AI learns something that in the end gives one of the actors a big advantage?

AI has many options for enhancing verification and detecting cheating, but also raises the verification problem to a higher level: Who will verify that the AI can be trusted (see the text by Maaik Verbruggen in this volume)? This question raises, first, the issue of the importance of “explainable AI” in the context of bi- or multilateral arms control and the related necessity of certification of training data by all participating states (Boulanin et al., 2020, p. 5). Second, it suggests that for the foreseeable future AI will at best assist arms controllers and verification experts rather than replace them. AI will only be implemented where mistakes and errors will do no significant harm: The IAEA, for example, has “been pursuing learning-based algorithms to help automate detecting and tracking events of interest to free up inspector time for more complex tasks” (IAEA, 2020, p. 10).

¹⁷<https://www.llnl.gov/news/researchers-developing-deep-learning-system-advance-nuclear-non-proliferation-analysis>, retrieved January 15, 2020.

The final problem is over-enthusiasm. Many experts are thrilled by what their AI prototypes can achieve and many offer deep insights into areas that were notoriously hard to look into in the past. However, as debated before, verification always walks a fine line between transparency and espionage and there are deliberate limitations that are accepted by everyone involved—the Open Skies Treaty, for example, which limits the resolution of observation photos. It is important to respect the limits of verification vis-à-vis intrusion. Experts must therefore always ask whether their AI is really helping enhance trust, is revealing secrets, or is even offering the possibility of misuse for spying.

5 Conclusion

As the significant increase in the publication of machine learning-related approaches to arms control problems show, reliability and practicability has reached a level where initial practical tests seem viable. At the lowest level, advances in AI-enhanced translation software make it easier for arms control experts to access written material or interact with entities being inspected. In addition, AI offers the chance to significantly enhance arms control measures as more data can be analyzed in a shorter time, freeing resources for the specifically tricky cases where human intelligence is needed. The fact that international organizations like the IAEA or the CTBTO have been holding workshops on the issue for some years now, or at least fostered debate, shows that ML-based AI will be of tremendous importance in the field in the future. It is no wonder that AI is either starting to be used exactly where the data situation is extensive and reliable (i.e., international organizations) or where data is open or easy to obtain and abundant, the realm of export controls and non-proliferation. As there are more and more toolkits available, it is becoming possible to develop and train AI models without much background knowledge, potentially leading to a boom of civil society arms controllers in the future—a societal Arms Control 2.0. As debated above, however, verification always has an underlying political dimension, and not every violation against the text of a treaty must be a violation of its spirit. False positives might lead to severe political disruptions. At least now, therefore, there still seems to be a continuous need for trained human inspectors and the “appropriate combination of algorithm-human fusion will be the subject of much analyst management effort in the coming years” (Steward et al., 2018, p. 23). A first important step would be to provide arms control experts with more fundamental information on artificial intelligence to facilitate communication. The cases debated in this text show that if arms controllers and AI experts work together, they can significantly improve verification and compliance, thereby taking the wind out of the sails of those who have written off verifiable arms control as an issue.

References

- Altmann, J. (2019). Confidence and security building measures for cyber forces. In C. Reuter (Ed.), *Information technology for peace and security* (pp. 185–203). Springer.
- Altmann, J. (2020). Advances in seismic and acoustic monitoring. In I. Niemeyer, M. Dreicer, & G. Stein (Eds.), *Nuclear non-proliferation and arms control verification* (pp. 231–248). Springer.
- Altmann, J., Linev, S., & Weiß, A. (2002). Acoustic-seismic detection and classification of military vehicles—Developing tools for disarmament and peace-keeping. *Applied Acoustics*, 63(10), 1085–1107.
- Bast, H., & Celikik, M. (2010). Efficient fuzzy search in large text collections. *ACM Transactions on Information Systems*, 9(4), Article 39. 49 pages.
- Bast, H., & Korzen, C. (2017). A benchmark and evaluation for text extraction from PDF. *ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017*, 1–10. <https://doi.org/10.1109/JCDL.2017.7991564>
- Baute, J. (2018). Introduction to the JNMM issue on open source and geospatial information analysis. *Journal of Nuclear Material Management*, XLVI(3), 4–7.
- Boulanin, V., Brockmann, K., & Richards, L. (2020). *Responsible artificial intelligence research and innovation for international peace and security*. Retrieved from https://www.sipri.org/sites/default/files/2020-11/sipri_report_responsible_artificial_intelligence_research_and_innovation_for_international_peace_and_security_2011.pdf
- Brase, J. M., McKinzie, E. G., & Zucca, J. J. (2020). Enhancing verification with high-performance computing and data analysis. In I. Niemeyer, M. Dreicer, & G. Stein (Eds.), *Nuclear non-proliferation and arms control verification* (pp. 299–307). Springer.
- Carter, A. (1989). *Success and failure in arms control negotiations*. Oxford University Press.
- Din, A. M. (Ed.). (1987). *Arms and artificial intelligence*. Oxford University Press.
- Feldman, Y., Arno, M., Carrano, C., Ng, B., & Chen, B. (2018). Toward a multimodal-deep learning retrieval system for monitoring nuclear proliferation activities. *Journal of Nuclear Material Management*, XLVI(3), 68–80.
- Gastelum, Z. N. (2020). Societal verification for nuclear nonproliferation and arms control. In I. Niemeyer, M. Dreicer, & G. Stein (Eds.), *Nuclear non-proliferation and arms control verification* (pp. 169–183). Springer.
- Gastelum, Z. N., Rutkowski, J., & Feldman, Y. (2018). A note from the editors of the special issue. *Journal of Nuclear Material Management*, XLVI(8–9).
- Gastelum, Z. N., & Shead, T. M. (2018). Inferring the operational status of nuclear facilities with convolutional neural networks to support international safeguard verification. *Journal of Nuclear Material Management*, XLVI(3), 37–47.
- Gayler, N. (1986). Verification, Compliance, and the Intelligence Process. In K. Tsipis, D. W. Hafemeister, & P. Janeway (Eds.), *Arms control verification. The technologies That make it possible* (pp. 3–13). Pergamon Brassey's.
- Goldblat, J. (1982). *Agreements for arms control: A critical survey*. Taylor & Francis.
- Gottenmeller, R. (2020). Foreword. In I. Niemeyer, M. Dreicer, & G. Stein (Eds.), *Nuclear nonproliferation and arms control verification* (pp. v–vi). Springer.
- Gubrud, M., & Altmann, J. (2013). *Compliance measures for an autonomous weapons convention*. ICRC Working Paper #2. Retrieved from https://www.icrac.net/wp-content/uploads/2018/04/Gubrud-Altman_Compliance-Measures-AWC_ICRAC-WP2.pdf
- Hochmuth, O., Keil, G., Winkler, F., Linev, S., & Meffert, B. (2001). *Mehrkanalige, hochauflösende Sensorstation für die Klassifikation schwerer Landfahrzeuge*. In Paper presented at the Frühjahrstagung der Deutschen Physikalischen Gesellschaft, Hamburg. Retrieved from <https://www3.informatik.hu-berlin.de/~hochmuth/bvp/dpg2001.pdf>
- IAEA. (2020). *Emerging technologies workshop. Insights and actionable ideas for key safeguard challenges*. Workshop Report STR-397. Retrieved from <https://www.iaea.org/sites/default/files/20/06/emerging-technologies-workshop-290120.pdf>

- Jasani, B., & Barnaby, F. (1984). *Verification technologies—The case for surveillance by consent*. Berg Publishers.
- Karkoszka, A. (1977). *Strategic disarmament, verification and national security*. Taylor & Francis for SIPRI.
- Keir, D., & Persbo, A. (2020). History, status and challenges for non-proliferation and arms control verification. In I. Niemeyer, M. Dreicer, & G. Stein (Eds.), *Nuclear non-proliferation and arms control verification* (pp. 15–26). Springer.
- Lück, N. (2019). *Machine learning powered artificial intelligence in arms control*. PRIF Report 8/2019. Retrieved from https://www.hsfk.de/fileadmin/HSFK/hsfk_publicationen/prif0819.pdf
- Molinier, M., Laaksonen, J., & Häme, T. (2007). Change detection (with self-organising maps). *Geo: Geoconnexion International Magazine*, 6(3), 36–38.
- Müller, H. (2017). *Learning unit 1: WMD, Conventional weapons and arms control: Basic concepts*. Retrieved from <https://nonproliferation-elearning.eu/learningunits/arms-control-basics/>
- Müller, H., & Schörnig, N. (2006). *Rüstungsdynamik und Rüstungskontrolle: Eine exemplarische Einführung in die Internationalen Beziehungen*. Nomos.
- Rogers, T. W., Jaccard, N., Morton, E. J., & Griffin, L. D. (2017). Automated x-ray image analysis for cargo security: Critical review and future promise. *Journal of X-Ray Science and Technology*, 25(1), 33–56.
- Russel, S., Vaidya, S., & Le Bras, R. (2010). *Machine learning for comprehensive nuclear-test-ban treaty monitoring*. CTBTO Spectrum 14. Retrieved from https://www.ctbto.org/fileadmin/user_upload/pdf/Spectrum/2010/Spectrum14_page32_machinelearning.pdf
- Rutkowski, J., Canty, M. J., & Nielsen, A. A. (2018). Site monitoring with sentinel-1 dual polarization SAR imagery using google earth engine. *Journal of Nuclear Material Management*, XLVI(3), 48–59.
- Schelling, T. C., & Halperin, M. H. (1961). *Strategy and arms control*. The Twentieth Century Fund.
- Schörnig, N. (2017). *Preserve past achievements! Why drones should stay within the missile technology control regime (for the time being)*. PRIF Report No. 149. Retrieved from https://www.hsfk.de/fileadmin/HSFK/hsfk_publicationen/prif149.pdf
- Schulze, J., Grüne, M., John, M., Neupert, U., & Dirk, T. (2020). Futures research as an opportunity for innovation in verification technologies. In I. Niemeyer, M. Dreicer, & G. Stein (Eds.), *Nuclear non-proliferation and arms control verification. Innovative systems concepts* (pp. 187–204). Springer.
- Silva, A., & Kligenboeck, M. (2019). *Enhancing safeguards verification work with innovative technology*. Retrieved from <https://www.iaea.org/newscenter/multimedia/videos/enhancing-safeguards-verification-work-with-innovative-technology>
- Steward, I., Lee, A., & ElGebaly, A. (2018). Automated processing of open source information for nonproliferation purposes. *Journal of Nuclear Material Management*, XLVI(3), 21–36.
- Sundaresan, L., Chandrashekar, S., & Jasani, B. (2017). Discriminating uranium and copper mills using satellite imagery. *Remote Sensing Applications: Society and Environment*, 5, 27–35.
- Versino, C., Cagno, S., Heine, P., & Carrera, J. (2018). A visual atlas on strategic trade. *Journal of Nuclear Material Management*, XLVI(3), 10–20.
- Williams, H. (2020). *Remaining relevant: Why the NPT must address emerging technologies*. Retrieved from London: <https://www.kcl.ac.uk/csss/assets/remaining-relevant-new-technologies.pdf>
- Withorne, J. (2020). *Machine learning applications in nonproliferation: Assessing algorithmic tools for strengthening strategic trade controls*. CNS NonPro Notes. Retrieved from <https://nonproliferation.org/machine-learning-applications-in-nonproliferation-assessing-algorithmic-tools-for-strengthening-strategic-trade-controls/>

Verifying the Prohibition of Chemical Weapons in a Digitalized World



Alexander Kelle and Jonathan E. Forman

● آقای هوش مصنوعی ●

🏢 رسانه هوش مصنوعی دانشگاه تهران 🏢

@MrArtificialintelligence

Abstract Kelle and Forman provide an analysis of the verification provisions of the Chemical Weapons Convention (CWC), which was negotiated in the 1980s and entered into force in 1997. Since then digitalization and the adoption of AI has progressed significantly. After introducing the convention's verification mechanisms, the authors discuss those treaty provisions as well as the latest Scientific Advisory Board report that deal with scientific and technological (S&T) advances of relevance to the Convention. Based on this, Kelle and Forman analyse various intersections of CWC verification and S&T advances. They conclude that the CWC has been drafted in a manner that allows it to adapt to S&T progress, thereby enabling its continued effective verification in an increasingly digitalized world.

1 Introduction

Chemical weapons (CW) comprise the entirety of toxic chemicals that are used to harm people or animals. They were widely used during World War I and in several instances since then, most recently in Malaysia, the United Kingdom, the Syrian Arab Republic and Russia (OPCW, 2017a, 2017b; Costanzi & Koblentz, 2019; Stone, 2020). The 1993 Chemical Weapons Convention (CWC) codifies the international community's agreement on the prohibition of chemical weapons. The CWC's robust and unique verification regime, which is implemented by the Organisation for the Prohibition of Chemical Weapons (OPCW), distinguishes it from other arms control and disarmament treaties and contributes to its successful implementation.

Former OPCW Science Policy Adviser, (2013–2020)

A. Kelle (✉)
IFSH, Berlin, Germany
e-mail: kelle@ifsh.de

J. E. Forman
The Hague, the Netherlands

The OPCW has verified the destruction of more than 98% of declared chemical weapons stockpiles. This is complemented by the CWC's industry verification regime, set up pursuant to Article VI of the Convention, which contributes to upholding the confidence of CWC states parties that chemicals are not diverted for purposes prohibited by the Convention.

This confidence was called into question by repeated reports of CW use in Syria starting in late 2012. In response to developments in Syria, the country's accession to the CWC in October 2013, and increasing evidence of an offensive Syrian CW program, in 2014 the OPCW's Technical Secretariat began to supplement its regular verification activities with non-routine missions with a view to clarifying all of the outstanding issues related to the Syrian initial declaration and to addressing the allegations of use of toxic chemicals as weapons in that country. The work of the Declaration Assessment Team (DAT) and the Fact-Finding Mission in Syria (FFM), both of which were established by the OPCW Director-General to uphold the object and purpose of the Convention, has demonstrated the OPCW's resilience in addressing unexpected situations and its capacity to adapt. While the dynamics of international relations and political conflicts have visibly challenged the work of the OPCW, necessitating this resilience, advances in science and technology (S&T) are also extensively cited as a challenge to the full and effective implementation of the CWC.

Science and technology continually evolve, and while in 1993 after the negotiations that led to the CWC had concluded, we may not have foreseen the specific S&T landscape before us in the 2020s, it should not be a surprise that much has changed. Recognizing this inevitability, the drafters of the CWC embedded mechanisms in the Convention's Article VIII to allow the OPCW to stay abreast of S&T developments of relevance to the Convention. The three main elements of these mechanisms are (1) the requirement for the OPCW to consider the use of S&T advances in its verification activities, (2) the mandate for the OPCW's Conference of States Parties (CSP) to review S&T as part of the quinquennial CWC review conferences, and (3) the establishment of a Scientific Advisory Board (SAB) to provide advice to the OPCW Director-General on S&T of relevance to the CWC.

The adoption of AI-based tools and digitalization is having a profound effect across scientific disciplines, and chemistry is no exception. It should be noted, however, that discussions of artificial intelligence and its bearing on the implementation of the CWC are a recent phenomenon, both in relation to the traditional verification activities performed by the OPCW and to the organization's more recently added investigative capabilities discussed below.

After a general description of the CWC verification provisions, the second part of the chapter outlines CWC provisions that deal with S&T advances of relevance to the Convention and presents relevant parts of the SAB report to the Fourth CWC Review Conference in November 2018. The third part draws the issues of verification and S&T advances together by discussing CWC verification in an increasingly digitalized world.

2 Verification Under the Chemical Weapons Convention

Effectively verifying the prohibition of chemical weapons under the CWC is based on the Convention itself and its Annex on Verification. In order to understand how new technologies including AI fit into this elaborate mechanism, to evaluate where they offer benefits, and to assess how difficult the implementation of new technologies would be, this section offers a basic understanding of the different verification rules and procedures implemented by the OPCW Technical Secretariat and CWC states parties.

In this context it is important to keep in mind that *declarations* are the basis of the CWC's verification regime; the accuracy, completeness and timelines of initial declarations by CW possessor states are essential for the functioning of the system. All measures described in the following section are guided by this basic idea.

2.1 Routine Verification

The aim of routine verification is to ensure that all states' declarations about either chemical weapons stockpiles (when relevant) or civilian industrial declarations match reality.

Routine verification activities related to chemical weapons as carried out by the OPCW Technical Secretariat are based on CWC Articles IV and V, and Parts IV and V of the CWC Verification Annex. These treaty provisions contain detailed declaration requirements for CW possessor states, and systematic on-site verification of storage and destruction activities (Trapp & Walker, 2014; Trapp, 2014a).

Routine industry verification activities for all member states with declarable industrial chemistry infrastructure are based on Article VI CWC and Parts VI to IX of the Verification Annex (Sossai, 2014). Here the OPCW's verification activities aim at confirming the absence of prohibited activities. States parties have to submit initial and annual declarations and have to accept data monitoring and on-site verification of facilities through the OPCW inspectorate. For verification purposes, the CWC distinguishes between four categories of chemicals, three of which are grouped into so-called "schedules" in the CWC's Annex on Chemicals. Chemicals are listed on these three schedules depending on the degree of risk they pose to the object and purpose of the Convention and on their use in the chemical industry (Trapp, 2014b). The fourth category consists of so-called "discrete organic chemicals" that could also pose a risk to the object and purpose of the CWC.

2.2 *Non-routine Verification*

The drafters of the CWC foresaw two types of non-routine verification missions that states parties could request should there be suspicion of non-compliance: a challenge inspection (CI) and an investigation of alleged use of CW (IAU). According to Article IX (8) “each State Party has the right to request an on-site challenge inspection of any facility or location in the territory or in any other place under the jurisdiction or control of any other State Party for the sole purpose of clarifying and resolving any questions concerning possible non-compliance with the provisions of this Convention” (Chemical Weapons Convention, 1997). Similarly, with respect to IAUs, Article X (8) stipulates that “each State Party has the right to request and, subject to the procedures set forth in paragraphs 9, 10 and 11, to receive assistance and protection against the use or threat of use of chemical weapons” (Chemical Weapons Convention, 1997).

It is noteworthy that since entry-into-force of the CWC in April 1997 neither a CI nor an IAU has ever been requested. However, other non-routine missions, most notably in relation to the Syrian CW program, have been established for:

- Clarification of Syria’s initial declaration, involving the establishment in 2014 of a Declaration Assessment Team; and
- Investigation of the alleged use of chemical weapons, for which a Fact-Finding Mission was created, also in 2014.

As the Syrian initial declaration raised several concerns, the DAT was set up. The team has undertaken over 20 visits to/consultations with Syria since 2014. In spite of these efforts, unresolved issues remain. As the OPCW Director-General has stated, the Technical Secretariat “is not able to resolve all identified gaps, inconsistencies and discrepancies in the declaration of the Syrian Arab Republic, and therefore cannot fully verify that Syria has submitted a declaration that can be considered accurate and complete in accordance with the Convention or Council decision EC-M-33/DEC.1” (OPCW, 2017a, 2017b). As this assessment remained unchanged as of spring 2020, the Director-General identified both FFM and declaration-related issues among the main areas of attention for the OPCW’s future operations in Syria (OPCW, 2020a).

The OPCW’s FFM was also established in April 2014 after continued reports of suspected CW use. Allegations of CW use in new incidents have continued to occur, with instances of sarin, sulfur mustard, and chlorine use as a weapon investigated by the FFM and the UN-OPCW Joint Investigative Mechanism (JIM). As the name of the FFM implies, it operates under a mandate to establish facts, not identify perpetrators. From 2015 to 2017 this latter role was assigned to the JIM, which issued several reports containing findings from its investigation. These findings considered a series of incidents and implicated actors on both sides of the conflict (Kelle, 2019).

2.3 Beyond Traditional Verification

In June 2018 OPCW member states convened the Fourth Special Session of the Conference of States Parties and agreed to establish a new investigation and identification team (IIT) in the OPCW. Identification adds another layer to traditional OPCW verification activities to effectively address instances of use of CW. Only with information specifying who the non-complying actor is can CWC States Parties and the OPCW policy making organs effectively address compliance matters pursuant to the provisions of the Convention (Kelle, 2019; OPCW, 2020b).

Given the wide range of expertise, technologies and methods that are both currently in use and under development in forensic science, S&T is a critical part of attribution work. Recognizing the valuable and actionable information that modern investigative techniques can provide, at its Twenty-Fourth Session the SAB recommended the establishment of a temporary working group (TWG) to conduct an in-depth review of methods and technologies that could be used by OPCW inspectors for investigative work. Capabilities enabled through these methods and technologies are crucial for the non-routine contingency operations that the Technical Secretariat has been increasingly deploying (OPCW-Scientific Advisory Board, 2018).

3 Science and Technology Under the CWC

3.1 CWC Provisions

The linkages between science and technology and the CWC are firmly embedded in the Convention itself. Article VIII (6) stipulates that “in undertaking its verification activities the Organization shall consider measures to make use of advances in science and technology.” Article VIII (21h) mandates the Conference of the States Parties to “review scientific and technological developments that could affect the operation of this Convention and, in this context, direct the Director-General to establish a Scientific Advisory Board to enable him [. . .] to render specialized advice in areas of science and technology relevant to this Convention.” Finally, Article VIII (22) calls on the quinquennial CWC Review Conferences to “take into account any relevant scientific and technological developments.”

3.2 The SAB Report to the 4th CWC Review Conference: Toward a Holistic Approach to Verification

As is customary during preparations for a Review Conference the SAB provided a report on relevant S&T developments to the Fourth Review Conference in

November 2018. This review of “Advances in Science and Technology” addresses inter alia developments in areas such as computational chemistry, Big Data and informatics, artificial intelligence, forensic science, remote sensing, and unmanned systems. The report goes on to recommend that both the SAB and the TS should continue to assess developments in these fields with relevance to the Convention (OPCW, 2018, p .5).

The SAB’s Temporary Working Group on verification, which provided significant input into the report, “considered opportunities arising from technological change for ensuring the Secretariat’s verification activities remain fit for purpose, and was of the view that particular attention should be given to remote/automated monitoring equipment, satellite imagery and information analysis tools.” (OPCW, 2018, p.44).

The report advised that “effective verification is not the assessment of an individual data point . . . , but rather all relevant data points pertaining to the site and State Party” and encouraged the Technical Secretariat “to move towards a comprehensive systems-based approach where all the separate elements of information are combined and analysed systematically” (OPCW, 2018, p. 45).

The SAB further emphasized that “effective use of data analysis, data mining, statistical analysis, and attribution analysis would serve to enhance existing capabilities for verification purposes.” Taking this further, the SAB recommended that the Secretariat put into place an information management structure that could provide the support required for the verification process . . . [leading to] a more analytical approach to verification, using all available information (declarations, inspection reports, satellite imagery, open source information, . . .).” (Ibid.)

Expanding beyond data analytics and the synthesis of diverse data streams to achieve a holistic picture of verification, the SAB also recommended that remote and/or automated monitoring technologies be added to the list of approved inspection equipment. (Ibid). These technologies add an intriguing dimension to the informatics capabilities considered by the Board, as they enable the collection of data that in principle could be fed into information analysis algorithms in real time.

More succinctly, through its focused considerations on verification and its broader S&T review, the SAB has recognized technological opportunities to enhance capabilities for verification of the Convention in both routine and non-routine contexts. These opportunities are addressed in the next section.

4 CWC Verification in the Age of Digitalization

4.1 Digitalization, Artificial Intelligence and Security

In a security context, AI receives significant attention for the new challenges, vulnerabilities and risks it presents. These concerns include military use and weaponization (Sisson et al., 2020) and extend well beyond Weapons of Mass Destruction (WMD) non-proliferation and disarmament (Cheatham et al., 2019). Across the

sciences, AI has been an enabler for new developments and advances (OECD, 2020). In biology, which has become as much a field of information and computational science as it is a natural science (Wintle et al., 2017), AI has been transformative, enabling capabilities for manipulating and understanding biological systems that were once limited to the realm of science fiction. The resultant changes in the life sciences have been referred to as a “new scientific revolution” (Chui et al., 2020) and have raised significant concerns about new capabilities for producing biological weapons using synthetic biology and gene editing, especially when combined with AI and other enabling technologies (Brockmann et al., 2019). The science of chemistry is also experiencing profound changes through adoption of AI-based tools and digitalization (Deloitte, 2017; World Economic Forum, 2020), although AI in the context of chemistry is less visible in WMD security discussions than in those on the life sciences. This is not to say that significant chemical security concerns have gone unnoticed. Frequently encountered examples in security-focused discussions include the use of AI tools to design novel toxic chemicals with effects on life processes that might not respond to state-of-the-art medical countermeasures or are perhaps more toxic than traditional chemical warfare agents, AI tools combined with automation for on-demand automated synthesis of toxic compounds, and especially the prospect of cyberattacks on the operating systems of chemical production facilities that might result in a large-scale chemical disaster (Stoye, 2015). Much has been researched and written on the security concerns and dangers of AI. The references cited in this section provide rich context on these issues.¹

One significant impact of digital transformation in the chemical enterprise is that chemical security cannot be fully separated from cybersecurity (Department of Homeland Security, 2015). In addition to the security concerns already touched upon, access to sensitive chemical information, which might include methods for producing chemicals of security concern, add to the risks that AI poses to chemical weapons non-proliferation.

We cannot ignore the security concerns about AI, its potential risk to chemical disarmament and non-proliferation, and the potential challenges we may need to be prepared to address. The security issues are significant and will only increase in complexity. It is also important to appreciate that AI and the digitalization that it enables has been steadily finding its way into and transforming a multitude of sectors and applications for more than a quarter of a century (Bughin et al., 2019). Digitalization is no longer an emerging trend. It drives and supports how we conduct business and operations and it continues to evolve (Deloitte, 2019, 2020), which challenges us to think of how it might also be used in beneficial ways so that chemical disarmament and non-proliferation are not left behind or placed at a significant disadvantage, while the world at large embraces the digital revolution.

¹The authors do not feel they can make meaningful additions to the wealth of reference material on the threat space within the confines of this chapter.

4.2 *What Is Artificial Intelligence?*

AI is a term that is used in a number of different ways. It is often used to describe software systems possessing “general” or “general-purpose” intelligence (Berruti et al., 2020) which, despite the progress made in developing AI systems, has yet to be realized (Benaich & Hogarth, 2020). There is also debate on what is actually possible (Fjelland, 2020).

The AI we currently see in use might be better thought of as software techniques (algorithms) that instruct computers to perform tasks. The techniques are often defined by terms such as “machine learning,” “probabilistic reasoning,” “fuzzy logic,” “logic programming” and “ontology engineering” (WIPO, 2019). These terms describe methods for performing computational tasks (WIPO, 2019). AI also does not represent a single technology (Benaich & Hogarth, 2020), but rather requires that the algorithms be integrated with computers and data in order to function (Buchanan, 2020), with additional technology possible needed for providing data.

From this integrated computer, algorithm, and data perspective, AI might be thought of as a component of a system which may require other components to perform operational tasks. These other components might be sensors to collect data, a camera to generate images that the AI can recognize, a vehicle to operate, or a game to play or combinations of various data and/or information-collecting components. It is often the case that such AI tools have limited versatility (Bergstein, 2020): they are built for specific tasks and coupled with components appropriate to successfully performing that task. Even while performing the task for which the AI has been designed, if the data used to train the AI has significant differences from the use case where the AI is deployed (Heaven, 2020a) or is exposed to situations that have not previously been encountered, the algorithms may “break”—ironically there are even examples of AI being affected by COVID19, with abrupt changes in human behavior confusing algorithms (Heaven, 2020b). The use of AI tools in healthcare has provided some illustrative case studies of how the types and characteristics of training data can lead to success (Hao, 2020) or failure (Heaven, 2020a).

AI tools can often outperform humans in many tasks, especially when it comes to processing and searching for information, playing games (especially for games where the best human players must think as many moves ahead of the current turn as they possibly can), and for tasks that require extended amounts of repetition.

4.3 *Digitalization in Routine Chemical Weapons Convention Verification*

The chemical industry, one of the Chemical Weapons Convention’s most important stakeholders, has been steadily increasing its adoption of AI-enabled digitalization as evidenced by the significant investment in “Industry 4.0” (Deloitte, 2017; Elser,

2019; Lin et al., 2019; Microsoft, 2019; World Economic Forum, 2020). Effectively this is chemical production enabled by smart supply chains and smart factories where equipment is augmented with connectivity and robotics to create fully integrated cyber-physical information systems. Industrial internet of things (IoT) sensors collect data and track information, with capabilities for visualizing the entire production chain and making decisions in real time. Ideally, all aspects of the chemical enterprise, including research and development, supply chains, manufacturing processes, sales and marketing, and customer support, are integrated into a system enabled by AI-driven big data analytics.

The adoption of Industry 4.0 by chemical companies improves operational efficiency, digitally enables innovative product offerings, accelerates innovation cycles, intensifies collaboration and data sharing across the value chain, develops new and more flexible business models, and improves customer interaction (Deloitte, 2017; Elser, 2019; Lin et al., 2019; Microsoft, 2019; World Economic Forum, 2020). Hidden under all this is something that directly affects treaty implementation—greater ability to track and report information, with the potential to streamline regulatory reporting and thus declarations under the CWC. While it is unrealistic to think that a CWC inspection would be given any access to industry data for verification purposes, the system does present the possibility of the use of data analytics for verification of declarations, e.g. “sampling and analysis” of data.

While the SAB reports do not explicitly state it, recommendations for holistic data-driven verification capabilities combined with remote and automated data collection tools necessitate a convergence of the virtual world of data and algorithms with the physical world of chemical inventories, supply chains, physical locations and biochemical traces indicative of the presence of, or exposure to, a chemical weapon agent. Realizing this convergence requires integration of a diversity of data streams (with both structured and unstructured data types) to establish connections between observable and measurable physical objects and entities. Conceptually, the Board would appear to be discussing the same basis upon which Industry 4.0 is being developed, only in the context of verification.

The SAB’s consideration of how AI and digitalization are opportunities which offer potential benefits to the verification system does not lie outside the mandates of the CWC. Paragraph 6 of the Convention’s Article VIII mandates the OPCW to consider the use of advances in science and technology in verification, which is exactly what the SAB is proposing. It is not that the SAB is unconcerned about security risks of AI or its potential misuse to aid in the realization of new chemical threats. Rather, in recognizing that the adoption and uses of these tools will only increase within the chemical enterprise, the Board is recommending that to keep pace with development in S&T relevant to the CWC these cannot be ignored. This recommendation includes consideration of the potential these technologies have when used for verification purposes.

The capability to integrate data streams and track and cross-check information has clear applications for verification of declarations submitted to the OPCW. The Technical Secretariat has already embarked on its own digital transformation in this area, beginning with the development of systems such as the Secure Information

Exchange (SIX) (OPCW, 2020c) and a recently rolled out Electronic Declaration Information System (EDIS) (OPCW, 2020d).

Industry 4.0 supply chain concepts have also been recognized as potentially valuable for implementation, where for example digital ledgers (like “Blockchains”) (Elser, 2019) might be drawn upon for diminishing and eliminating discrepancies between the quantities of scheduled chemicals declared by Member States reporting transfers. The Blockchain as a digital record of all transactions of a given chemical product (from its manufacture to the end user) provides a mechanism for verifying reported (or for recognizing and correcting incorrectly reported) imports and exports of scheduled chemicals. The life cycle of chemical products (from raw materials to production to disposal and/or recycling, and all steps in between) involves transfers between numerous entities (companies, transportation, customers, and more). A digitalized record of transactions could not only provide a means for more effectively flagging transfer discrepancies, but could also act as a chain of custody. In such an application, it might be more difficult to circumvent declarations and regulatory reporting. With digitalized systems giving rise to increased concerns about the use of technology to hide information, these types of decentralized approaches to tracking and reporting may warrant consideration.

The recognition of scheduled chemicals is another area where advanced informatics tools could provide great benefit. Outside the perspectives of chemists whose work focuses on the universe of atoms and molecules contained within the CWC’s Schedules, it may not be appreciated how all-encompassing these schedules are in regard to the types of chemicals they cover—there is an unlimited number of possible chemical compounds that could exist which meet the criteria established within these schedules (Timperley et al., 2018). The 34,254 chemicals in the OPCW’s Scheduled Chemicals Database (OPCW, 2020e) represent the scheduled chemicals that have been reported, but this is a mere fraction of what might exist (Pontes et al., 2020). Chemists who work with such compounds can quickly recognize previously unreported chemicals that meet the criteria for inclusion in a schedule from the molecular structure. However, officials at border stations or regulatory bodies who are not trained in chemistry or have no access to trained chemists may not realize a chemical is a scheduled chemical when it is not labeled with the exact name or number found on reference lists that they have access to (on which previously unreported scheduled chemicals are unlikely to be found). Digitalized tools that can match chemical structures to schedule criteria would be beneficial for end users who do not speak the “language of chemistry” (OPCW, 2019b).

Chemists who work with scheduled chemicals could also benefit from AI tools capable of recognizing and assigning molecular information to corresponding schedules. In test methods for scheduled chemicals, the presence of a specific chemical is confirmed by comparison of its analytical data (most commonly using mass spectra) to validated reference data. The reference data is obtained by analysis of an actual sample of the chemical being verified. AI tools that recognize characteristic signals (“peaks”) in mass spectral data which correlate with specific molecular structural features open up potential for detection of scheduled chemicals without a previously

obtained reference data set. Such methods have not been fully developed and, more importantly, validated for this application, although current research in this direction has produced encouraging results (Lim et al., 2018).

4.4 Digitalization in Non-routine Chemical Weapons Convention Verification

Up to this point we have touched on areas where digitalized tools and AI are being used or have potential for use in routine verification activities. The SAB and its TWG on investigative science and technology have given the most consideration to applications in non-routine situations.

Investigative work might use digital tools and technologies to collect verifiable information that can ensure the information is unaltered from its original form (and contains verifiable time stamps, geolocation information and other metadata). While this may not be specifically an AI tool, digitalized tracking (in the spirit of Industry 4.0 supply chains) could make such tools more powerful for investigative purposes. The SAB's TWG on investigative science and technology also recognized the value of advanced data analytics for a variety of other investigative purposes (OPCW, 2019a).

The use of AI tools to help analyze chemical sample data was previously mentioned in the context of test methods and proficiency testing. In the case of complex chemical samples (for example, where the chemicals of interest might be highly degraded, it may be difficult to obtain a background/control sample and/or the chemical of interest may have been fully metabolized by plants, animals or microbes), AI tools that can model and predict chemical uptake in plants (Bagheri et al., 2020) and metabolic products (Toubiana et al., 2019) might help in the detection of chemicals contained in samples that can verify a specific chemical exposure for investigative purposes.

AI tools also have an interesting potential for helping to recognize “unusual” events and phenomena. Areas where this has been demonstrated include agricultural applications, where a variety of remote sensing data streams are collected on crops which provide farmers with real-time actionable information on plant health (which allows immediate action such as watering, or applying fertilizer or pesticide to specific sections of a field) (Liakos et al., 2018).

Tools developed for agriculture include cellphone apps that can diagnose “plant diseases” from digital photos (based on machine-learning algorithms) (Mohanty et al., 2016). There is also significant interest in AI tools that can help to quickly determine and diagnose the health status of humans—this could be from combinations of images, measurements (some non-invasive such as temperature, pulse and other readings that can be taken on the spot by a medical responder) and biomedical test results (Yu et al., 2018). For this kind of application, it is noteworthy that chemical warfare agent toxidrome recognition from social media has been

demonstrated by forensic toxicologists (Toprak et al., 2020). While this was not accomplished by means of an AI tool, developing such a method is not inconceivable. AI tools that can help recognize signs and symptoms of toxic chemical exposure not only have investigative value but may also be useful in helping to quickly diagnose exposure and initiate treatment. We should not expect to see AI tools replacing medical professionals, but their ability to recognize chemically-induced symptoms that may not be commonly seen or immediately recognized by medical responders with limited experience with toxic chemical exposures could help to improve emergency or retrospective diagnosis of a chemical toxidrome.

AI tools may also present other opportunities for verification. Recognizing unusual chemical exposures in vegetation (whether predictive of a specific chemical or simply distinguishable from common presentation of unhealthy plants) may be a useful tool for investigative work (and could help identify samples for chemical analysis). The use of “precision agriculture”-derived techniques to recognize signs of chemical warfare agent exposure has actually been considered by the SAB (Forman et al., 2018), and some research groups (Kuska et al., 2018).

4.5 Will Digitalization Disrupt Verification Procedures?

We have highlighted a number of areas where AI tools provide potential benefits in chemical disarmament and non-proliferation, including several examples currently being used or researched. As AI continues to become more ubiquitous, keeping pace with evolving S&T will only demand its adoption and use to a greater degree. In this regard we note that such adoption is not absent from OPCW. Nonetheless, even with all the advances in S&T, the use of enabling technologies must continue to be looked at from a practical point of view. Field testing and validation of results are needed, and AI tools that can be relied upon for use in decision-making require appropriate datasets for training and validation—in some cases it may never be possible to acquire such data. Similarly, while not the focus of this paper, the security concerns of AI and digitalization cannot be discounted. These need to be understood and vulnerabilities identified in evaluating whether any new tools are suitable for an intended application. Similarly, cybersecurity capabilities, which need to be dynamic and constantly evolving, must be fit for purpose to counter the use of AI tools designed to circumvent verification or compromise digitalized verification tools. Uncertainties as to how much these systems can be trusted for use in verification and how resilient they are to exploitation of vulnerabilities will certainly challenge their adoption.

None of the issues described above are unique to considerations of the use of AI tools in treaty implementation, as demonstrated by on-going debate and discussion across sectors and policies where AI tools are used (Brundage et al., 2020). Nevertheless, we live in exciting times, the enabling power of AI is in principle only limited by our imagination and the data available for training algorithms. As the SAB continues its consideration of new tools and emerging technologies (Forman

et al., 2018; OPCW, 2019a), we look forward to seeing how these recommendations are discussed at the OPCW and in its policy-making organs.

5 Conclusions

In summary, AI and digitalization have evolved into key components of keeping pace with S&T, and we find it intriguing that the CWC mandates consideration of S&T advances for verification. AI and digitalization cannot be treated like a singular new technology that we monitor and somehow mitigate. These tools are embedded in our daily lives in ways we may not even be aware of. Many potential benefits are clear, but suitability and reliability for any given task require validation, as well as consideration of the often murky risks and security concerns of cyberspace. Independent of these concerns, we are already seeing interest in and dialogue concerning AI tools in chemical non-proliferation (Borrett et al., 2020). The SAB's advice has been forward-thinking in regard to this area of S&T, providing much information for others to build upon.

The drafters of the CWC did not intend that this Convention would only eliminate the chemical weapons of today. They also saw it as a tool for protecting and strengthening the norms against chemical weapons in perpetuity. As the CWC is underpinned and impacted by S&T, treaty implementation, and especially verification, keeping both abreast of, and keeping pace with the twenty-first century's evolving and dynamic S&T landscape is necessary for achieving this goal. The discussion of S&T in the context of security often appears to involve fear—at this juncture in time, with the rapid changes in S&T that we are seeing, we cannot afford to be afraid. This does not mean that we should ignore the risks and challenges, but rather that we must bring scientific literacy into our discussions and understand both possible and practical aspects of S&T advances—in both beneficial and malicious uses. In the case of AI and digitalization, these tools and technologies are not going to go away, and their enabling capabilities provide intriguing opportunities for treaty implementation. The SAB, in drawing upon opportunities and potential uses of emerging technology, provides a forward-looking perspective on supporting the CWC as we move into a future, which is sure to present new and unexpected challenges. Success in meeting these challenges will require new and innovative approaches. This is where S&T and the way it is adopted and implemented provides opportunity.

References

- Bagheri, M., Al-jabery, K., Wunsch, D., & Burken, J. G. (2020). Examining plant uptake and translocation of emerging contaminants using machine learning: Implications to food security. *Science of The Total Environment*, 698, 133999.

- Benaich, N., & Hogarth, I. (2020). *State of AI report 2020*. Retrieved December 10, 2020, from <https://www.stateof.ai/>
- Bergstein, B. (2020, February 19). What AI still can't do. *MIT Technology Review*. Retrieved from <https://www.technologyreview.com/2020/02/19/868178/what-ai-still-cant-do/>
- Berruti, F., Nel, P., & Whiteman, R. (2020, April 29). An executive primer on artificial general intelligence. *McKinsey Insights*. Retrieved from <https://www.mckinsey.com/business-functions/operations/our-insights/an-executive-primer-on-artificial-general-intelligence>
- Borrett, V., Hanham, M., Jeremias, G., Forman, J., Revill, J., Borrie, J., et al. (2020). *Science and technology for WMD compliance monitoring and investigations*. WMD Compliance and Enforcement Series no. 11. UNIDIR.
- Briggs, B., & Buchholz, S. (2019, January 16). *Tech trends 2019*. Deloitte. Retrieved from <https://www2.deloitte.com/us/en/insights/focus/tech-trends/2019/executive-summary.html>
- Brockmann, K., Bauer, S., & Boulanin, V. (2019). *Bio plus X: Arms control and the convergence of biology and emerging technologies*. Report. Stockholm International Peace Research Institute. Retrieved from <https://www.sipri.org/publications/2019/other-publications/bio-plus-x-arms-control-and-convergence-biology-and-emerging-technologies>
- Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., et al. (2020). *Toward trustworthy AI development: Mechanisms for supporting verifiable claims*. Retrieved from <https://arxiv.org/abs/2004.07213>
- Buchanan, B. (2020, August). *The AI triad and what it means for national security strategy*. Report. Center for Security and Emerging Technology. Retrieved from <https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Triad-Report.pdf>
- Buchholz, S., & Briggs, B. (2019, January 15). *Tech trends 2020*. Deloitte. Retrieved from <https://www2.deloitte.com/us/en/insights/focus/tech-trends/2020/tech-trends-introduction.html>
- Bughin, J., Manyika, J., & Catlin, T. (2019). *Twenty-five years of digitization: Ten insights into how to play it right*. McKinsey Global Institute. Retrieved from <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/twenty-five-years-of-digitization-ten-insights-into-how-to-play-it-right>
- Cheatham, B., Javanmardian, K., & Samandari, H. (2019, April 26). Confronting the risks of artificial intelligence. *McKinsey Quarterly*. Retrieved from <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence>
- Chemical Weapons Convention. (1997, April 29). *Convention on the prohibition of the development, production, stockpiling and use of chemical weapons and on their destruction*. Retrieved from <https://www.opcw.org/chemical-weapons-convention>
- Chui, M., Evers, M., & Zheng A. (2020, May 7). How the bio revolution could transform the competitive landscape. *McKinsey Quarterly*. Retrieved from <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/how-the-bio-revolution-could-transform-the-competitive-landscape>
- Costanzi, S., & Koblentz, G. (2019). Controlling Novichoks after Salisbury: Revising the Chemical Weapons Convention schedules. *The Nonproliferation Review*, 26(5-6), 599–612. <https://doi.org/10.1080/10736700.2019.1662618>
- Deloitte. (2017). *Chemistry 4.0 growth through innovation in a transforming world*. Retrieved from <https://www2.deloitte.com/global/en/pages/consumer-industrial-products/articles/cip-chemistry.html>
- Department of Homeland Security. (2015). *Chemical sector cybersecurity framework implementation guidance*. Retrieved from <https://www.cisa.gov/sites/default/files/publications/chemical-cybersecurity-framework-implementation-guide-2015-508.pdf>
- Elser, B. (2019). *AI & blockchain: Chemical industry insights and actions*. Accenture. Retrieved from <https://www.accenture.com/us-en/insights/chemicals/ai-blockchain-chemical-industry>
- Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, 7(1), 2–9. Retrieved from <https://www.nature.com/articles/s41599-020-0494-4>

- Forman, J. E., Timperley, M. C., Aas, P., Abdollahi, M., Alonso, I. P., Baulig, A., et al. (2018). Innovative technologies for chemical security. *Pure and Applied Chemistry*, 90(10), 1527–1557.
- Hao, K. (2020, October 2). How an AI tool for fighting hospital deaths actually worked in the real world. *MIT Technology Review*. Retrieved from <https://www.technologyreview.com/2020/10/02/1009267/ai-reduced-hospital-deaths-in-the-real-world/>
- Heaven, W. D. (2020a). Google’s medical AI was super accurate in a lab. Real life was a different story. *MIT Technology Review*. Retrieved from <https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/>
- Heaven, W. D. (2020b). Our weird behavior during the pandemic is messing with AI models. *MIT Technology Review*. Retrieved from <https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/>
- Kelle, A. (2019). The international regime prohibiting chemical weapons and its evolution. In N. Hynek, O. Ditych, & V. Stritecky (Eds.), *Regulating global security. Insights from conventional and unconventional regimes* (pp. 115–141). Palgrave Macmillan.
- Kuska, M. T., Behmann, J., & Mahlein, A. K. (2018). Potential of hyperspectral imaging to detect and identify the impact of chemical warfare compounds on plant tissue. *Pure and Applied Chemistry*, 90(10), 1615–1624.
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8), 2674.
- Lim, J., Wong, J., Wong, X. M., Tan, E. L., Chieu, H. L., Choo, D., & Neo, N. K. (2018). Chemical structure elucidation from mass spectrometry by matching substructures. *physics.chem-ph*: arXiv:1811.07886.
- Lin, S., Mulayath, S., Womack, D., & Zaheer, A. (2019). *Shift to enterprise-grade AI: How chemicals and petroleum leaders are adopting artificial intelligence*. Report. IBM. Retrieved from <https://www.ibm.com/thought-leadership/institute-business-value/report/chemicals-petroleum-ai#>
- Microsoft. (2019). *2019 manufacturing trends report*. Microsoft Dynamics 365. Retrieved from <https://info.microsoft.com/rs/157-GQE-382/images/EN-US-CNTNT-Report-2019-Manufacturing-Trends.pdf>
- Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7(1419). <https://doi.org/10.3389/fpls.2016.01419>
- OPCW-Scientific Advisory Board. (2018). *Summary of the first meeting of the Scientific Advisory Board’s temporary working group on investigative science and technology*. Document SAB-27/WP.1. OPCW. Retrieved from www.opcw.org/sites/default/files/documents/SAB/en/sab-27-wp01_e_.pdf
- Organization for Economic and Cooperation and Development. (2020). *The digitalization of science, technology and innovation: Key developments and policies*. Retrieved from https://www.oecd-ilibrary.org/science-and-technology/the-digitalisation-of-science-technology-and-innovation_b9e4a2c0-en
- Organization for the Prohibition of Chemical Weapons. (2017a, March 7). *Statement of the permanent representative of Malaysia to the 84th session of executive council—Report on the use of a chemical weapon in the death of a DRPK national*. Retrieved from https://www.opcw.org/fileadmin/OPCW/EC/84/en/Malaysia_ec84_statement.pdf
- Organization for the Prohibition of Chemical Weapons. (2017b). *Note by the director-general: Outcome of further consultations with the Syrian Arab Republic regarding its chemical weapons declaration*. Document EC-86/DG.30. OPCW.
- Organization for the Prohibition of Chemical Weapons. (2018). *Report of the Scientific Advisory Board on developments in science and technology for the fourth special session to review the operation of the chemical weapons convention*. RC-4/DG.1. OPCW. Retrieved from www.opcw.org/sites/default/files/documents/CSP/RC-4/en/rc4dg01_e_.pdf
- Organization for the Prohibition of Chemical Weapons. (2019a). *Investigative science and technology report of the Scientific Advisory Board’s temporary working group*. Document

- SAB/Rep/1/19. OPCW. Retrieved from <https://www.opcw.org/sites/default/files/documents/2020/11/TWG%20Investigative%20Science%20Final%20Report%20-%20January%202020%20%281%29.pdf>
- Organization for the Prohibition of Chemical Weapons. (2019b). *What is the language of chemistry?* Retrieved from www.opcw.org/sites/default/files/documents/2019/10/The%20Language%20of%20Chemistry-V2.pdf
- Organization for the Prohibition of Chemical Weapons. (2020a). *Note by the director-general: Progress in the elimination of the Syrian chemical weapons programme*. Document EC-96/DG.3. OPCW. Retrieved from www.opcw.org/sites/default/files/documents/2020/11/ec96dg03%28e%29.pdf.
- Organization for the Prohibition of Chemical Weapons. (2020b). *Note by the technical secretariat. First report by the OPCW investigation and identification team pursuant to paragraph 10 of decision C-SS-4/DEC.3 "Addressing the threat from chemical weapons use" Ltamenah (Syrian Arab Republic) 24, 25, and 30 March 2017* (document S/1867/2020). OPCW.
- Organization for the Prohibition of Chemical Weapons. (2020c). *Secure information exchange (SIX)*. Retrieved December 12, 2020, from www.opcw.org/resources/declarations/secure-information-exchange-six
- Organization for the Prohibition of Chemical Weapons. (2020d). *Electronic declaration information system (EDIS)*. Retrieved December 12, 2020, from www.opcw.org/resources/declarations/electronic-declaration-information-system-edis
- Organization for the Prohibition of Chemical Weapons. (2020e). *Scheduled chemicals database*. Retrieved December 12, 2020, from <https://apps.opcw.org/CAS/default.aspx>
- Pontes, G., Schneider, J., Brud, P., Benderitter, L., Fourie, B., Tang, C., et al. (2020). Nomenclature, chemical abstracts service numbers, isomer enumeration, ring strain, and stereochemistry: What does any of this have to do with an international chemical disarmament and nonproliferation treaty? *Journal of Chemical Education*, 97(7), 1715–1730.
- Sisson, M., Spindel, J., Scharre, P., & Kozyulin, V. (2020). *The Militarization of Artificial Intelligence*. United Nations Office of Disarmament Affairs.
- Sossai, M. (2014). Article VI: Activities not prohibited under the convention. In W. Krutzsch, E. Myjer, & R. Trapp (Eds.), *The chemical weapons convention. A Commentary* (pp. 173–194). Oxford University Press.
- Stone, R. (2020, September 8). How German military scientists likely identified the nerve agent used to attack Alexei Navalny. *Science*. doi: <https://doi.org/10.1126/science.abe6561>
- Stoye, E. (2015, June 10). Security experts warn chemical plants are vulnerable to cyber-attacks. *Chemistry World*. Retrieved from <https://www.chemistryworld.com/news/security-experts-warn-chemical-plants-are-vulnerable-to-cyber-attacks-/8632.article>
- Timperley, C. M., Forman, J. E., Abdollahi, M., Al-Amri, A. S., Alonso, I. P., Baulig, A., et al. (2018). Advice on chemical weapons sample stability and storage provided by the Scientific Advisory Board of the Organisation for the Prohibition of Chemical Weapons to increase investigative capabilities worldwide. *Talanta*, 188(1), 808–832.
- Toprak, S., Can, E. Y., Altinsoy, B., Hart, J., Dogan, Z., & Ozcetin, M. (2020). Social media video analysis methodology for sarin exposure. *Forensic Sciences Research*. <https://doi.org/10.1080/20961790.2020.1825061>
- Toubiana, D., Puzis, R., Wen, L., Sikron, N., Kurmanbayeva, A., Soltabayeva, A., et al. (2019). Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. *Communications Biology*, 2(214).
- Trapp, R. (2014a). Article V: Chemical weapons production facilities. In W. Krutzsch, E. Myjer, & R. Trapp (Eds.), *The chemical weapons convention. A commentary* (pp. 151–172). Oxford University Press.
- Trapp, R. (2014b). Annex on chemicals. In W. Krutzsch, E. Myjer, & R. Trapp (Eds.), *The chemical weapons convention. A commentary* (pp. 173–194). Oxford University Press.

- Trapp, R., & Walker, P. (2014). Article IV: Chemical weapons. In W. Krutzsch, E. Myjer, & R. Trapp (Eds.), *The chemical weapons convention. A commentary* (pp. 173–194). Oxford University Press.
- Wintle, B. C., Boehm, C. R., Rhodes, C., Molloy, J. C., Millett, P., Adam, L., et al. (2017). Point of View: Transatlantic perspective on 20 emerging issues in biological engineering. *eLife*. Retrieved from <https://elifesciences.org/articles/30247#abstract>
- WIPO. (2019). *WIPO technology trends 2019: Artificial intelligence*. World Intellectual Property Organization. <https://www.wipo.int/publications/en/details.jsp?id=4386>
- World Economic Forum. (2020). *Digital transformation: Chemistry & advanced materials industry*. Retrieved from <https://reports.weforum.org/digital-transformation/chemistry-advanced-materials/>
- Yu, K., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2, 719–731.

AI and Biological Weapons



Filippa Lentzos 

● آقای هوش مصنوعی ●

🏠 رسانه هوش مصنوعی دانشگاه تهران 🏠

@MrArtificialintelligence

Abstract Lentzos highlights key impacts of machine learning and automation on biological research, medicine and healthcare, emphasizing how these developments could make the production of biological weapons easier and proliferation more likely. While biological weapons are completely prohibited by the Biological Weapons Convention, artificial intelligence and other converging technologies are radically transforming the dual-use nature of biology and present significant challenges for the treaty. The chapter discusses these challenges and presents a vision for how biological arms control can evolve to remain relevant in the Fourth Industrial Revolution.

1 Introduction

A “game changer” that is radically transforming the dual-use nature of biology is how Eleonore Pauwels, an internationally renowned expert and director of the Anticipatory Intelligence Lab at the Woodrow Wilson International Center for Scholars, characterized artificial intelligence (AI) and its impact on biological weapons and arms control (Pauwels, 2019a). She was speaking to Biological Weapons Convention (BWC) delegates at their summer 2019 meeting to review science and technology developments, having been invited to share her views on emerging technologies relevant to the 1972 multilateral treaty. It was the first time AI was given serious consideration at the treaty’s annual meetings. Pulling no punches, Pauwels explained that “the convergence of biotechnologies with cyber and AI technologies covers the whole spectrum of the bioeconomy, from precision medicine to infectious disease to the bioindustry. It will deeply impact how long we live, how we treat illnesses, and our view of our place on the biological continuum,” also

This chapter draws on my earlier article ‘How to protect the world from ultra-targeted biological weapons’ published in *The Bulletin of Atomic Scientists* Vol.76(6):302–308.

F. Lentzos (✉)
King’s College London, London, UK
e-mail: filippa.lentzos@kcl.ac.uk

adding “yet, beyond the promises, we also face an era of hybrid security threats that are poorly understood, leaving the multilateral system unequipped to anticipate and prevent emerging risks” (Pauwels, 2019a).

The BWC is the principal legal instrument banning biological warfare and the deliberate use of bacteria and viruses to inflict harm. The treaty itself is relatively short, comprising only 15 articles, but over the years, the treaty articles have been supplemented by a series of additional understandings reached at treaty review conferences. The term “verification,” traditionally thought of as the foundation of post-Second World War weapons treaty compliance regimes, does not feature in the text of the BWC (Lentzos, 2019a). Efforts in the 1990s to develop a verification mechanism for the treaty failed, and the main role and responsibility for BWC compliance continues to fall on the treaty’s 183 states parties. The Security Council is to act as the final arbitrator on allegations of compliance breaches, but it has not to date been requested to investigate any allegations. The World Health Organization, the Food and Agriculture Organization of the United Nations and the World Organization for Animal Health have potential roles in clarifying ambiguous events and situations, but they only provide expert information to help states parties, and this does not amount to determining compliance.

This chapter draws together the surprisingly scant consideration given by the biological arms control community to date to the potentially transformational power of AI for biology. It provides an overview of some of the main impacts of machine learning and automation on biological research, medicine and healthcare expected by experts, emphasizing how these developments could make the production of biological weapons easier and proliferation more likely. The chapter then turns to discussing how AI could help in furthering the aim of the BWC to ensure biological weapons are not developed or used. In closing, the chapter considers how best biological arms control can evolve to reflect the radical transformations AI is already beginning to introduce into the biological field.

2 Adding Computing Power to Bioinformatics

Genomic technologies are driving a vast expansion in genomic data, from gene sequences and entire genomes to data that links genes to specific functions and other types of metadata for humans, other animals, plants and microbes. This data is becoming increasingly digitized (Bajema et al., 2018), and computational power is significantly changing how genomic data is analyzed. This integration of AI computation into biology opens up new possibilities for understanding how genetic differences shape the development of living organisms, including ourselves, and how these differences make us and the rest of the living world susceptible to diseases and disorders, and responsive to drugs and treatments.

Advanced pattern recognition and abstracting statistical relationships from data—the hallmarks of machine and deep learning—have shown significant potential to help researchers make sense of complex genomic datasets and extract clinically

relevant findings. Take two prominent examples: functional genomics and tailored drug discovery.

The ability of machine learning to link, correlate and analyze data is particularly useful for interpreting gene functions and identifying genetic markers responsible for certain diseases (Brockmann et al., 2019). Known as functional genomics, this field of research makes it possible to predict how likely someone is to develop diseases such as type 1 diabetes or breast cancer or to develop certain traits and capabilities, such as someone's height or resistance to specific pathogens that result from complex genetic influences. Deep learning also enables computer-based experimentation for functional genomics, and work is underway to predict how genetic sequences might function before they are assembled—even if the combination has not been observed in nature. Computational power has also helped researchers understand the evolving relationship between our genotypes, phenotypes (physical characteristics) and microbiomes (the bacteria and viruses that live on and inside the human body), as well as to improve our genotype-phenotype functional knowledge of pathogens.

Private industry has been instrumental in developing data-mining techniques. Google's genomic AI platform, DeepVariant, for example, has been at the forefront of developing an automated, deep-learning approach to identifying genetic variants in an individual genome from billions of short sequences (Poplin et al., 2018).

Adding computational power to drug development facilitates “parallel read operations of 10 billion nucleic acid molecules in a single experiment” and can increase experimental precision down to single-molecule manipulations (Spiez Laboratory, 2018, p. 34). The use of deep learning to develop new drug candidates has overcome many limitations of physics-based models, enabling models to be built from simple representations of chemical and biological entities and automating suggestions of synthesizable structures with improved properties. It also highlights how the convergence of automation with evolutionary algorithms vastly expands the number of materials that can be synthesized, tested and optimized.

In developing new drug candidates, a robot can reportedly screen over 10,000 compounds per day through conventional “brute-force” tactics (University of Cambridge, 2015). However, while simple to automate, this approach is still relatively slow and wasteful because every compound in the library has to be tested. The first AI robot to automate early-stage drug design came on stream in 2015 (Williams et al., 2015). Called “Eve,” it was developed by researchers at the Universities of Aberystwyth and Cambridge, who had earlier developed “Adam,” a machine to independently discover scientific knowledge. To make screening processes for potential drugs intelligent, Eve randomly selects a subset of compounds from a library, carries out various tests on them, and, based on the compounds that pass the tests, uses statistics and machine learning to predict new structures that might achieve an even better score (University of Cambridge, 2015).

Private companies also contribute substantially to the development of machine learning in drug development. The pharmaceutical giant Novartis, for example, used computational power to develop a vaccine in less than 3 months from the first reported cases of humans becoming infected with H7N9 influenza virus (IAP,

2015, p. 11). In another example, Deep Genomics uses its AI platform to map pathological genetic pathways in identifying drug candidates.

Advances in function genomics and drug discovery, as well as in other areas, offer the possibility of developing bespoke, or personalized, treatments using machine learning analysis of genomic and health data. “Precision public health” aims to deliver the right intervention to the right population at the right time, and it is already beginning to deliver genomic-based interventions for health and health care (Khoury et al., 2018). The Centers for Disease Control and Prevention promotes its wide use of artificial intelligence and machine learning to improve public health surveillance (such as forecasting of influenza) and disease detection, mitigation and elimination (Siordia & Khoury, 2020). While still in its early days, precision medicine—spanning personalized vaccines and antibodies, personalized treatment relying on virology and microbe research, personalized cancer treatments and treatments involving in vivo gene editing—is also starting to become a reality (Regalado, 2018).

A number of private companies, such as Tempus, IBM and Pfizer, are actively exploring possibilities, though these efforts are mostly focused on understanding how machine learning could help identify genetic markers or patients that should or could be candidates for personalized treatments (Pauwels & Vidyarthi, 2017). Experts emphasize that there is “still pervasive uncertainty about how accurate deep machine-learning will be in drawing useful inferences between the different datasets that make our biology” (Pauwels & Vidyarthi, 2017, p. 5).

3 Mounting Security Concerns

Various risk assessment frameworks have been used to get a sense of the potential security risks arising from the mix of artificial intelligence and biotechnology (O’Brien & Nelson, 2019). But balancing the more general reach needed to capture a broad scope of converging technologies in the life sciences with the need to maintain enough specificity to capture nuances has proven difficult. The main security concerns boil down to worries that, if the intent were there, the convergence of emerging technologies could be used to speed up the identification of harmful genes or DNA sequences (Brockmann et al., 2019). More specifically, there are concerns that adding advanced pattern recognition to genomic data could significantly facilitate: the enhancement of pathogens to make them more dangerous; the modification of low-risk pathogens to become high-impact; the engineering of entirely new pathogens; or even the re-creation of extinct, high-impact pathogens like the variola virus that causes smallpox. These possibilities are arising at a time when new delivery mechanisms for transporting pathogens into human bodies are also being developed. In addition to the bombs, missiles, cluster bombs, sprayers and injection devices of past biowarfare programs, it could now also be possible to use drones, nano-robots or even insects.

Added to these pathogen-specific risks are traditional cyber risks and “cyberbiosecurity” risks focused particularly on the bioeconomy (Murch & DiEuliis, 2019). Cyberbiosecurity risks include waging adversarial attacks on automated biocomputing systems, biotech supply chains or strategic cyberbiosecurity infrastructure. Malicious actors could, for example, use AI malware to co-opt networks of sensors and impact control decisions on biotech supply chains with the intent to damage, destroy or contaminate vital stocks of vaccines, antibiotics and cell or immune therapies. In another scenario, AI malware could be used to automate data manipulation with the intent to falsify, erase or steal intelligence within large curations of genomics data. Such data poisoning could affect how pathogens are detected and analyzed. It could also affect bio-intelligence on complex diseases in subpopulations collected over many years.

The merger of the biological data revolution with computing power has created another serious security concern: ultra-targeted biological warfare. In past biowarfare programs, weapons targeted their intended victims through geographic location. Advances in biotechnology open up the possibility that malicious actors could deploy a biological weapon over a broad geographic area but only affect targeted groups of people, or even individuals.

The possibility of such “genetic weapons” was first discussed in the biological arms control community in the 1990s, as the Human Genome Project to map the full complement of human genes got underway. The UK government said, “it cannot be ruled out that information from such genetic research could be considered for the design of weapons targeted against specific ethnic or racial groups” (BWC/CONF. IV/4). The British Medical Association cautioned that “the differential susceptibility of different populations to various diseases” had been considered in the past, and that “whilst we should hope that genetic weapons are never developed, it would be a great mistake to assume that they never can be, and therefore that we can safely afford to ignore them as a future possibility” (BMA, 1999). A report from the Stockholm International Peace Research Institute (SIPRI) spoke of the potential for “future development of weapons of mass extermination which could be used for genocide” (SIPRI, 1993).

Developments in genomic technologies and other emerging technologies, especially machine and deep learning, have spurred renewed concerns. “Access to millions of human genomes—often with directly associated clinical data—means that bio-informaticists can begin to map infection susceptibilities in specific populations,” a recent report from the United Nations Institute for Disarmament Research warned (Warmbrod et al., 2020). A United Nations University report, meanwhile, asserts that “deep learning may lead to the identification of ‘precision maladies,’ which are the genetic functions that code for vulnerabilities and inter-connections between the immune system and microbiome (Pauwels, 2019b). Using this form of bio-intelligence, malicious actors could engineer pathogens that are tailored to target mechanisms critical in the immune system or the microbiome of specific subpopulations.” A 2018 National Academies of Sciences report suggests “[a]ctors may consider designing a bioweapon to target particular subpopulations based on their genes or prior exposure to vaccines, or even seek to suppress the

immune system of victims to ‘prime’ a population for a subsequent attack (NASEM, 2018). These capabilities, which were feared decades ago but never reached any plausible capability, may be made increasingly feasible by the widespread availability of health and genomic data.”

It is important to note that there are barriers limiting access to targeted biological weapons. The technical base, expertise and funding required for the design of a targeted biological weapon suggest that only a significantly resourceful and motivated actor would be likely to explore this possibility (Lentzos, 2017).

4 How Should the BWC Respond?

Ultra-targeted biological weapons are relatively unlikely to be used because of the complexity required to create them (Brockmann et al., 2019). If the purpose is to harm a specific individual or group, most malevolent actors would probably resort to more low-tech or direct methods, such as firearms or poison. Unfortunately, this is not a sufficient basis for biological arms control in the twenty-first century. As one of the great champions of biological disarmament, Matthew Meselson, professor of molecular biology at Harvard University, reflected in 2000 as he contemplated the century ahead in an essay entitled “Averting the Hostile Exploitation of Biotechnology”: “[A]s our ability to modify fundamental life processes continues its rapid advance, we will be able not only to devise additional ways to destroy life but will also become able to manipulate it—including the processes of cognition, development, reproduction and inheritance. . . . Therein could lie unprecedented opportunities for violence, coercion, repression, or subjugation” (Meselson, 2000).

The current BWC regime comprehensively prohibits biological weapons, understood as biological agents used for harmful purposes. Parties to the treaty agree that the BWC unequivocally covers all microbial or other biological agents or toxins, naturally or artificially created or altered, as well as their components, whatever their origin or method of production.

On the whole, this covers the pathogen-specific risks and risks of ultra-targeted weapons. Indeed, the UK government, which first raised the issue of genetic weapons as a possibility in the mid-1990s, specifically stated that genetic weapons would be a “clear contravention” of the treaty (BWC/CONF.IV/4). Cyberbiosecurity risks are not covered by the BWC, but the BWC and arms control treaties more generally are not appropriate instruments to address these sorts of risks.

Where there might be some uncertainty is where harm does not involve biological agents (Lentzos & Invernizzi, 2018). Developments in science and technology are making novel biological weapons conceivable that, instead of using bacteria or viruses to make us sick, directly target the immune, nervous or endocrine systems, the microbiome, or even the genome by interfering with, or manipulating, biological processes. This could be achieved, for example, by using a construct based on synthetic structures created or inspired by DNA or RNA, but not qualifying as DNA, RNA or any other known, naturally occurring nucleic acid. In this sort of

case, the coverage of the BWC is less clear, but the intent of the treaty to prohibit such harm is beyond doubt.

The real challenge for the treaty, however, is not in its coverage but in ensuring states parties comply with it and live up to their obligations. This is particularly difficult as relevant materials, equipment and technical know-how are diffused across multiple and varied scientific disciplines and sectors. Moreover, biological agents themselves exist in nature and are living organisms generally capable of natural reproduction and replication.

The dual-use nature of biology and the challenges it poses for compliance assessment was already recognized in the early phase of BWC treaty negotiations. In a 1968 statement to the predecessor of the Conference on Disarmament, the United Kingdom noted, for instance, that “no verification is possible in the sense of the term as we normally use it in disarmament discussions” (Mulley, 1968). In other words, it was not considered possible to verify the BWC with the same level of accuracy and reliability as the verification of nuclear treaties such as the Treaty on the Non-Proliferation of Nuclear Weapons (NPT), which was negotiated immediately prior to the BWC. Consequently, Article I of the BWC—through which states “agree to never under any circumstances acquire or retain biological weapons”—is vague in demarcating the borders of prohibited and legitimate activities. Article I merely refers to biological agents “of types and in quantities that have no justification for prophylactic, protective or other peaceful purposes.”

This “general purpose criterion” of the BWC means the treaty permits almost any kind of research for defensive or protective purposes. Some of this work is justifiable. Other research treads closer to the blurred line between defensive and offensive work. The trouble for states in distinguishing permitted biodefense projects from prohibited projects is that it is not possible to assess the facilities, equipment, material and activities involved alone, but the purpose, or intent, of those activities must also be examined and interpreted (Lentzos & Littlewood, 2018).

The significant, and accelerating, advances in the ability to manipulate genes and biological systems, alongside developments in emerging technologies such as AI, automation and robotics, and the rise in biodefense programs and build-up in capacities (Koblentz & Lentzos, 2016) mean that Cold War-era tools of compliance assessment are becoming increasingly outdated.

To determine if there is intent, it is not enough to simply count fermenters, measure the sizes of autoclaves, and limit amounts of growth media. More and more states recognize that biology, to a large extent, defies material accountancy-type verification methodologies. The United Kingdom, for instance, recently noted that BWC compliance is “much more one of transparency, insight and candour, rather than material balances or counting discrete objects such as fermenters” (United Kingdom, 2019).

Somewhat ironically in our ever-expanding digital world, a shift is underway from quantitative approaches and binary models of compliance assessment in biological arms control to more qualitative methods. Leading states are exploring means of demonstrating good practices and responsible science through new voluntary initiatives that enable them to demonstrate transparency and build trust;

initiatives such as peer review, implementation review and transparency visits (Lentzos, 2019b). These information-sharing initiatives emphasize interaction and flexibility, expert-level exchanges of best practices rather than just on-site monitoring, and a broad conception of relevant laboratories and facilities—and they have been deemed to add real value to compliance judgments by participating states (Belgium et al., 2016).

Similarly, as a way to complement laws and regulations around biosecurity, civil society groups have led the development of norm-building controls such as codes of conduct, prizes, awards, competitions and other incentives for good behavior. The flip-side—leveraging reputational risks, corporate shaming and social pressures for poor behavior—is also beginning to be explored. It has become abundantly clear that, in the Fourth Industrial Revolution, compliance with the BWC needs to be less about verifying a binary state—being “in compliance” or “not in compliance”—and more about analyzing justifications provided for the activities in question and managing dual-use potential.

5 Evolving Biological Arms Control

For biological arms control and the UN more generally, the broader challenges involve extending the management regime to stakeholders other than countries, particularly to private industry and civil society groups, but also to other entities, and maintaining contemporary relevance as the global forum for security debates on emerging technologies. Technologies, like biotechnologies, that have traditionally been compartmentalized in elite siloed institutions and national labs and monitored by national governments are now increasingly accessible to and even controlled by private tech platforms and research communities around the world. In the era of AI, limiting access to intangible transfer of knowledge and tools involving dual-use research will only become more difficult.

There is a narrowing window of opportunity to evolve biological arms control on a structured basis. One way to do this—and to link the biological field with other emerging technologies—is to actively encourage collaborations across AI, cyber and biotechnologies in order to develop responsible security practices where scientists learn enough about each converging field and its impact on dual-use research. Yet, on its own, this type of collaboration is not enough to protect the world from misuse of powerful and converging technologies.

One idea for improving the management of broad and fast-paced technological advances involves a World Economic Forum-like “network of influence,” composed of exceptional individuals from business, academia, politics, defence, civil society and international organizations, to act as a global “Board of Trustees” to oversee developments relevant to biological threats in science, business, defense and politics and to decide on concerted cross-sector actions (Lentzos, 2019c). A similar idea—not limited to the biological field but cutting across emerging technologies—would develop a “Global Foresight Observatory” comprising a constellation of key public

and private sector stakeholders convened by a strategic foresight team within the UN (Pauwels, 2019b). The Board of Trustees or the Global Foresight Observatory could be supplemented by a secondary oversight layer that enlists individuals and select institutions to act as “sentinels” (Lentzos, 2019c). These sentinels could have dual functions: first, to actively promote responsible science and innovation, and second, to identify security risk for consideration by the Board or the Observatory.

These new governance structures could be supplemented by political initiatives—AI and bioinformatics groups, for example—to establish a new type of transparency, confidence-building and BWC compliance assessment, and to support the prevention of biological weapons development and the management of dual-use biological research. The colossal challenges of converging technologies will require bold ideas like these to re-envision future biological arms control.

Acknowledgements Warm thanks to Cédric Invernizzi and Alex Lampalzer for their input on the chapter and for their always constructive discussions on biological arms control. The chapter also benefitted from discussions with colleagues at a set of CIFAR-hosted workshops on AI, arms control and international security, held in Toronto and Santa Monica in 2019 and 2020.

References

- Bajema, N. E., DiEuliis, D., Lutes, C., & Lim, J. B. (2018). *The digitization of biology: Understanding the new risks and implications for governance*. Research Paper No.3. Center for the Study of Weapons of Mass Destruction. National Defense University.
- Belgium, Canada, Chile, Czech Republic, France, Ghana, Germany, Luxembourg, Mexico, the Netherlands, Spain, Switzerland & the United States. (2016, November 10). *Building confidence through voluntary transparency exercises*. BWC/CONF.VIII/WP/35.
- British Medical Association (BMA). (1999). *Biotechnology, weapons and humanity*. CRC Press.
- Brockmann, K., Bauer, S., & Boulanin, V. (2019, March). *Bio plus X: Arms control and the convergence of biology and emerging technologies*. Stockholm International Peace Research Institute.
- BWC/CONF.IV/4. (1969, October 30). *Background paper on new scientific and technological developments relevant to the Convention on the prohibition of the development, production and stockpiling of bacteriological (biological) and toxin weapons and on their destruction*. [https://docs-library.unoda.org/Biological_Weapons_Convention_-_Fourth_Review_Conference_\(1996\)/BWC_CONF.IV_04.pdf](https://docs-library.unoda.org/Biological_Weapons_Convention_-_Fourth_Review_Conference_(1996)/BWC_CONF.IV_04.pdf)
- IAP: The Global Network of Science Academies. (2015). *The biological and toxin weapons convention: Implications of advances in science and technology*. <https://www.interacademies.org/30337/The-Biological-and-Toxin-Weapons-Convention-Implications-of-advances-in-science-and-technology>
- Khoury, M. J., Engelgau, M., Chambers, D. A., & Mensah, G. A. (2018). Beyond public health genomics: Can big data and predictive analytics deliver precision public health? *Public Health Genomics*, 21(5–6), 244–250.
- Koblentz, G. D., & Lentzos, F. (2016). *21st century biodefence: Risks, trade-offs & responsible science*. BWC Review Conference Series Paper No.3. International Law & Policy Institute. <http://nwp.ilpi.org/wp-content/uploads/2016/11/03-21st-century-biodefence-gold.pdf>
- Lentzos, F. (2017, July 3). Ignore bill gates: Where bioweapons focus really belongs. *Bulletin of Atomic Scientists*. <https://thebulletin.org/2017/07/ignore-bill-gates-where-bioweapons-focus-really-belongs/>

- Lentzos, F. (2019a). *Compliance and enforcement in the biological weapons regime*. Paper 4. WMD Compliance and Enforcement Series. United Nations Institute for Disarmament Research.
- Lentzos, F. (2019b). Trust and transparency in biodefense. In S. K. Singh & J. H. Kuhn (Eds.), *Defense against biological attack*. Springer.
- Lentzos, F. (2019c). Re-thinking biological arms control for the 21st Century. *Fletcher Security Review*, 6(1), 33–36.
- Lentzos, F., & Invernizzi, C. (2018, January 5). DNA origami: Unfolding risk? *Bulletin of Atomic Scientists*. <https://thebulletin.org/2018/01/dna-origami-unfolding-risk/>
- Lentzos, F., & Littlewood, J. (2018, July 28). DARPA's Prepare program: Preparing for what? *The Bulletin of Atomic Scientists*. <https://thebulletin.org/2018/07/darpas-prepare-program-preparing-for-what>
- Meselson, M. (2000). Averting the hostile exploitation of biotechnology. *Chemical and Biological Conventions Bulletin*, 48, 16.
- Mulley, F. W. (1968, August 6). *Statement by Frederick W. Mulley, Minister of State at the Foreign Office, United Kingdom*. Eighteen-Nation Disarmament Committee, ENDC/ PV.387.
- Murch, R., & DiEuliis, D. (2019). Editorial: Mapping the cyberbiosecurity enterprise. *Frontiers in Bioengineering and Biotechnology*, 7, 235.
- National Academies of Sciences, Engineering and Medicine (NASEM). (2018). *Biodefense in the age of synthetic*. Biology, National Academies Press.
- O'Brien, J. T., & Nelson, C. (2019). Assessing the risks posed by the convergence of artificial intelligence and biotechnology. *Health Security*, 18(3), 219–227.
- Pauwels, E. (2019a, August 2). *Cyber-AI-Bio convergence presentation to BWC meeting of experts 2* [video]. UN Geneva. <http://webtv.un.org/search/mx2-1st-meeting-biological-weapons-convention-meetings-of-experts-2019/6066151799001/?term=BWC%20meetings%20of%20experts%202019&page=3>. Presentation starts at around 12 minutes
- Pauwels, E. (2019b, April 29). *The new geopolitics of converging risks: The UN and prevention in the era of AI*. United Nations University Centre for Policy Research.
- Pauwels, E., & Vidyarthi, A. (2017). *Who will own the secrets in our genes? A U.S.–China race in artificial intelligence and genomics*. Wilson Briefs, Wilson Center.
- Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10), 983–987.
- Regalado, A. (2018). Look how far precision medicine has come. *MIT Technology Review*.
- Siordia, C., & Khoury, M. J. (2020). *Artificial intelligence, public trust and public health*. *CDC blog on genomics and precision health*. <https://blogs.cdc.gov/genomics/2020/09/17/artificial-intelligence/>
- SIPRI. (1993). *SIPRI yearbook 1993: World armaments and disarmament*. Oxford University Press.
- Spiez Laboratory. (2018). *Spiez convergence: Report on the third workshop*. Swiss Federal Office for Civil Protection.
- United Kingdom. (2019, July 10). *Institutional strengthening of the Convention: Reflections on the 2001 Protocol and the verification challenge*. Working Paper to the BWC Meeting of States Parties, BWC/MSP/2019MX.5/WP.1.
- University of Cambridge. (2015, February 4). *Artificially-intelligent robot scientist 'Eve' could boost search for new drug*. Press release. <https://www.cam.ac.uk/research/news/artificially-intelligent-robot-scientist-eve-could-boost-search-for-new-drugs>
- Warmbrod, L., Revill, J., & Connell, N. (2020). *Advances in science and technology in the life sciences*. United Nations Institute for Disarmament Affairs (UNIDIR).
- Williams, K., Bilsland, E., Sparkes, A., Aubrey, W., Young, M., Soldatova, L. N., et al. (2015). Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *Journal of the Royal Society Interface*, 12(104), 20141289.

Doomsday Machines? Nukes, Nuclear Verification and Artificial Intelligence



Jana Baldus 

● آقای هوش مصنوعی ●

🏠 رسانه هوش مصنوعی دانشگاه تهران 🏠

@MrArtificialintelligence

Abstract This chapter aims to paint a clearer picture of the use of AI and autonomy in nuclear weapon systems. It asks how AI and autonomy are (and have been) used thus far: Do qualitative improvements make nuclear weapons more secure or more unreliable, unstable and unpredictable instead? Do technological advances increase the risks of nuclear weapons use or can they help prevent nuclear war? The answer is far from obvious. The same technology that could increase advance warning time could lead to more complex decision cycles and shorter reaction times in crises. AI could contribute to preventing the spread of nuclear weapons, help enhance verification systems, and increase transparency among states. On the other hand, the use of AI also creates its own verification problems. Accordingly, this chapter aims to provide a balanced assessment of the possible benefits and dangers of integrating AI or autonomy in nuclear weapons and nuclear verification systems.

1 Introduction

In the movie series *The Terminator* an artificial superintelligence called *Skynet* triggers a nuclear war among humans, causing the death of three billion people. Out of fear of being shut down, *Skynet* takes control of US nuclear forces and launches a nuclear attack against Russia (which blindly retaliates). This futuristic scenario seems to be the pop-cultural yardstick by which the artificial intelligence-nuclear weapons nexus is measured. The good news is: if our greatest fear is a revengeful, self-conscious superintelligence that wages nuclear war against human-kind, we have nothing to worry about (at least for now): “The state of the art, while impressive, still trails a long way behind the cultural perception of what autonomous systems ought to be able to do in a military context, namely operate safely and reliably in [a] complex, uncertain and adversarial environment” (Boulanin, 2019).

J. Baldus (✉)

Peace Research Institute Frankfurt, Frankfurt, Germany

e-mail: baldus@hsfk.de

The tale of benefits and dangers of new technologies seems to be at its strongest in the nuclear field. The use of artificial intelligence (AI) and autonomy holds the promise of strategic advances for nuclear warfare: improved early warning and threat detection systems, optimization of logistics, even enhanced decision-support systems (Horowitz, 2019; Sayler, 2019; Sauer, 2019). On the other hand, nuclear-weapon states fear that they could lose their second-strike capability if they fail to keep pace with technological progress. The biggest players in the nuclear field are already outperforming each other with more and more capable systems (for an overview, see Boulanin et al., 2020). Russia gave a glimpse of the possible future when it announced the development of *Poseidon*, a nuclear-powered, nuclear-armed underwater drone in 2015.¹ The US on the other hand is developing a new stealth bomber—the B-21 Raider—that is capable of carrying both conventional and nuclear weapons and will only optionally be crewed (Gertler, 2019).² These two examples illustrate how autonomy can creep into nuclear weapon systems. And indeed, some media are already warning of worst-case scenarios such as “catastrophic nuclear use” or even what has been dubbed “nuclear holocaust,” as in the science fiction world of *The Terminator*. But the tale about new and enhanced weapons technologies only shrouds one fairly simple fact: in general terms, AI (in a narrow sense) is already being used in nuclear weapon systems and has been for a long time.

The aim of this chapter is to paint a balanced picture of the use of AI and autonomy in nuclear weapon systems. It asks how AI and autonomy are (and have been) used in nuclear weapon systems thus far and what implications their use may have. But the chapter also addresses the question of which future applications of AI and what levels of autonomy in those systems are conceivable or considered likely. Artificial intelligence is referred to here in its narrow application. It is defined as 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” (Boulanin, 2019). As a subfield of AI, machine learning (ML) is understood as “an approach to software development that first builds systems that can learn and then teaches them what to do using a variety of methods” (ibid.). However, machine learning cannot be compared with the human cognitive process. Instead, ML-based systems learn by finding statistical relationships in data. The biggest advantage of learning machines is thus that, once developed, they reduce the need for human programming (depending on how capable the systems are and which approach to learning they use) (Horowitz et al., 2019; Scharre and Horowitz 2018). Autonomy, finally, is understood as the “ability of a machine to execute a task, or tasks, without human input, using interactions computer programming with the environment” (Boulanin & Verbruggen, 2017). AI, and in particular ML, are enablers for increasing autonomy. Without the potential of AI (such as increased

¹ Allegedly, Russia commenced underwater tests in 2018 or early 2019 (Gady, 2019a, 2019b).

² Russia also announced a new generation of strategic bombers (to be introduced by 2040), which will be uncrewed (Atherton, 2020).

processing power, improvement of sensors and image recognition) to enhance machine perception, autonomy would not be possible. Autonomy can thus be regarded as a “by-product” (Boulanin, 2019) of scientific advances in AI.

The best approach appears to be to assess the possible benefits and dangers of integrating AI or autonomy in nuclear weapon systems with a degree of skeptical caution. Do qualitative improvements make nuclear weapons more secure, or more unreliable, unstable and unpredictable instead? Do technological advances increase the risks of nuclear weapons use or can they help prevent nuclear war? Given the inherent ambivalence of new technologies, it is difficult to give a simple answer to these questions. The same technology that could increase advance warning time could lead to more complex decision cycles and shorter reaction times in crises. On the one hand, the use of AI could contribute to the development of more effective verification systems for monitoring nuclear nonproliferation and disarmament, but it also creates its own verification problems. These questions require in-depth research, which is, however, still in its infancy. To date, the limited scientific literature available mainly focuses on possible impacts of new technologies on nuclear stability, on the current or future capabilities of nuclear weapon systems, and on potential risks that are associated with the use of AI or autonomy in nuclear weapon systems. Yet, this approach in addressing AI and autonomy ignores possible contributions AI or autonomy can make to nuclear arms control, especially with regard to efforts to promote nonproliferation and disarmament. AI could well extend the range of verification instruments available for nuclear arms control, disarmament and nonproliferation.

When navigating the field of nuclear weapons and new technologies, assumptions are frequently made which will be examined in more detail below:

1. The use of AI and autonomy in nuclear weapon systems is not a new phenomenon. Applications of early AI and (semi)automatic systems were already deployed during the Cold War (Borrie, 2019).
2. Advances in AI and increasing integration of AI and autonomy in nuclear weapon systems can cut both ways. They can potentially lead to disruptive changes in military operations and warfare. Then again, AI could, if handled carefully, make nuclear weapon systems safer and more dependable by minimizing the risk of human failure (Boulanin, 2019).
3. The use of AI or autonomy in conventional weapon systems can be potentially more problematic and destabilizing than applications of AI in nuclear weapons; it could reduce strategic stability by challenging second-strike capabilities and lowering the threshold for the use of nuclear weapons (Horowitz et al., 2019).
4. The possible disruptive influence of AI and autonomy on nuclear warfare depends as much on technological factors as on psychological factors, such as confidence in new technologies or one’s own capabilities as well as perceptions of an opponent’s capabilities and intentions (Geist & Lohn, 2018).
5. The increasing capabilities of AI could contribute to the development of new and more reliable instruments for nuclear arms control and verification. However, the effectiveness of such systems relies on political will; AI-enhanced verification

alone will not be insurance against treaty violations (Kaspersen & King, 2019; Lück, 2019).

2 AI in Nuclear Weapon Systems

Historically, both the US and the USSR used “first-wave” AI—such as so-called “expert” systems³ and automation to gain the upper hand in the Cold War and to ensure the survivability and thus retaliatory capability of their nuclear forces (Horowitz et al., 2019). Even then considerable effort was made to ensure that nuclear weapon systems and nuclear command and control would not operate without human supervision. Nuclear decision-makers in the US and USSR were “deeply aware that, when dealing with something as tightly-coupled, complex and potentially hazardous as nuclear command and control, machine-based systems face real limits that require meaningful human control and supervision” (Borrie, 2019). Automation was used in communication systems to enable automated broadcast networks to transmit emergency action messages in the event of nuclear attacks or for automated retargeting (Horowitz et al., 2019). Early AI was also used to develop and maintain the capability of nuclear powers to respond to nuclear attacks. Semiautomated ‘dead-hand’ systems (such as the Soviet Union’s “Perimeter” system⁴) should only be activated in exceptional cases when a decapitating attack on nuclear command and control occurred (Klare, 2020; Boulanin et al., 2020). Automation proved to be dangerous even though nuclear command and control largely remained under human supervision during the Cold War. In 1983 a Soviet early warning system reported incoming US intercontinental ballistic missiles with “highest” confidence. The system triggered an automated alert and call for retaliation—but the launch order still had to be activated by a human operator. The Soviet Lieutenant Colonel in charge, Stanislav Petrov, however, reported a system malfunction and ignored the counter-attack order, thereby preventing a nuclear confrontation (Topychkanov, 2019). The Stanislav-Petrov incident illustrates the potential destructive power of automation when paired with nuclear weapons. It also explains why “nuclear-armed states have historically limited the role of automation in nuclear launch platforms so that humans retain positive control over nuclear targeting and strike initiation” (Horowitz et al., 2019). Premature reliance on unstable or untested technology for the most destructive weapons ever invented could indeed lead to

³Expert systems are automated and/or semiautomated systems that follow a solely rule-based decision-making logic (if-then) (Horowitz et al., 2019). They differ from autonomous systems inasmuch as they do not have the capability to develop “and select from different courses of action to accomplish goals based on [their] knowledge and understanding of the world” (Boulanin & Verbruggen, 2017).

⁴The Soviet/Russian Perimeter system is said to be still (or rather back) in operation (Topychkanov, 2019).

horror scenarios such as falsely launched nuclear missiles, with the plausible consequence of nuclear responses and counterresponses.

2.1 Qualitative Improvements in Nuclear Weapon Systems

The Stanislav-Petrov incident exemplifies the dark side of (semi)automated nuclear command and control and the two sides of the coin inherent in using AI in nuclear weapon systems. This section aims to draw a clear(er) picture of how the use of AI or autonomy in nuclear weapon systems can change the field for better or worse. The perils and benefits associated with the use of AI or autonomy in nuclear weapons generally depend on the area of application: the use of AI could qualitatively improve threat detection and early warning systems, make nuclear command, control and communication (NC3) more reliable, and lead to greater accuracy in missile and guidance systems. In terms of nuclear strategy, the use of autonomy in nuclear launch platforms and delivery systems holds the promise of gaining new military advantages. On the other hand, even more than in the conventional field, it is essential that AI applications in nuclear weapon systems function without error, as any miscalculations would have greater repercussions than with conventional weapons. The two-edged sword of emerging technologies can cut deeper when coupled with the unique destructiveness of nuclear weapons.

AI could enhance detection and sensory capabilities, including those of early warning systems, thus reducing the risks of false positives. AI and, in particular, machine learning (ML) could accelerate the identification of threats in complex environments. What is more, AI and ML may give early warning systems greater perceptual intelligence for identifying signals or objects and situations of interest, such as mobile nuclear launchers or unusual troop movements (Cuihong, 2019; Horowitz et al., 2019). By facilitating the processing and analysis of big sets of data AI could help improve predictive models of the production, commissioning, deployment or use of nuclear weapons (Boulanin, 2019). The same applies to intelligence, surveillance and reconnaissance (ISR) data (Verbruggen, 2020). AI and increasing autonomy could allow the analysis of “large swaths of data quickly for anomalous behavior at a scale and speed that would not be possible with human analysts” (Horowitz et al., 2019). The use of machine learning to train early warning systems with additional data could lead to greater situational awareness and further reduce risks of false alarms, accidents or accidental use (Cuihong, 2019; Verbruggen, 2020). In the case of indicators of incoming nuclear strikes, increased system speed resulting from the inclusion of AI and a greater degree of autonomy in early warning or detection systems (i.e., faster response times) could, in the future, leave more room for reaction and for interpretation of data, as well as time to check for false positives (Klare, 2020). This is of particular importance in the nuclear realm, as the inadvertent use of nuclear weapons due to false alarms or time pressure is one of the most likely catalysts for nuclear war.

The inclusion of AI could not only help reduce risks of miscalculation and inadvertent escalation, but also help strengthen the protection of NC3 systems, increase combat readiness and optimize resource management (Boulanin, 2018; Gompert & Libicki, 2019). In terms of communication, AI and particularly autonomy could improve the ability of NC3 systems to transmit decisions faster and more efficiently and thus allow for more careful and rigid execution of launch orders (Horowitz et al., 2019). In the future, AI and increased autonomy could also help maintain communications in unstable environments or hazardous situations (Horowitz et al., 2019). This is most pertinent with regard to navigation and control of nuclear-capable submarines but could also extend to missions in degraded environments such as after nuclear weapons use. Finally, AI could boost qualitative improvements in nuclear delivery systems, in particular with regard to targeting. Image processing enhanced by ML could accelerate (and improve) the detection of relevant targets, their exact location and most vulnerable parts and thus optimize how warheads are allocated (Stefanovich, 2020).⁵ While this could give attackers a strategic advantage, it would also increase the speed of escalation and lower the threshold for nuclear use.

More potent than these qualitative improvements of nuclear delivery systems could be a greater degree of autonomy of nuclear launch platforms and delivery vehicles. What is increasingly normal in the conventional area is about to follow in the nuclear field: weaponized uncrewed aerial or underwater vehicles (UAV/UUV) which would then, potentially, be nuclear-armed (such as Russia's Poseidon or the US's new Stealth Bomber).⁶ Some strategic advantages of greater autonomy of nuclear launch platforms and delivery vehicles (i.e. the ability to autonomously launch, select and attack targets, but also to autonomously engage in ISR operations) are quite obvious: it would allow for extended endurance, greater reach and use in hardly accessible areas and contested (air-)spaces, greater persistence, greater mass (e.g. through swarming), and the recoverability of weapon systems (Boulanin, 2019). Yet, autonomy extends beyond the notion of nuclear "killer robots." As the Cold War has shown, automated or largely autonomous strategic response systems can be appealing to nuclear powers—particularly, when their deterrent value is in doubt. Dead-hand systems would enable a nuclear response even if NC3 were interrupted after a possible nuclear strike; autonomy would give dead-hand systems more perceptual intelligence to detect, analyze, and react to such situations. Some experts assert that a high degree of autonomy in NC3 could, in the future, secure communications even in the most degraded environments (e.g., Boulanin et al.,

⁵Image recognition through ML has made great progress in recent years but is still unreliable and can easily be exploited (Sayler, 2019). Until these deficiencies are overcome, it is doubtful whether the military would rely on systems supported by image recognition such as automated target recognition, at least when it comes to nuclear weapons.

⁶This development has long been anticipated: The Missile Technology Control Regime (MTCR), established in 1987, has treated drones restrictively, primarily because they could be potential candidates as delivery systems for nuclear weapons and other weapons of mass destruction (Schörnig, 2018).

2020; Horowitz et al., 2019). However, it would also make NC3 systems more susceptible to external interference or internal malfunctions.

At present it seems out of the question that an autonomous Skynet-like superintelligent system with final authority (and independent decision-making power) over nuclear weapons will ever be developed. Yet, even at a lower level, decision-makers seem to be hesitant to allow a greater degree of autonomy in nuclear weapon systems, at least when it comes to nuclear command and control (Freedberg, 2019).⁷ This includes the use of UAVs or UUVs as launch platforms for nuclear warheads—the ultimate decision-making power will remain with human operators for now, or to put it in everyday language: there will always be a “human-in-the-loop.” The picture is different when it comes to protecting critical infrastructure, for example in terms of cyber defense or the physical protection of launch platforms, or conducting difficult ISR missions (Boulanin et al., 2020). Here, decision-makers may be considerably more willing to include autonomy.

2.2 *The Other (Dark) Side of the Coin*

Most risks that are associated with the use of AI or autonomy in nuclear weapon systems are inherent qualities of emerging technologies: opacity, brittleness and external vulnerability, as well as possible biases in data sets that could affect the reliability of nuclear weapon systems (Horowitz et al., 2019). The use of algorithms could render systems more opaque to their operators. The resulting *trust gap* (i.e., the preference for human decision-making over technology that is seen as unreliable) could lead to a failure of human-machine interaction (ibid.). It is sometimes suggested that the destructiveness of nuclear weapons induces natural skepticism about a high degree of technologization. However, the opposite of the trust gap (i.e., the *automation bias*) could be equally problematic in cases where human operators display blind faith in the supposedly superior capabilities of a machine or algorithm, treating “the AI system’s suggestions as on par with or better than those of human advisers” (Geist & Lohn, 2018). This could be particularly dangerous with regard to early warning or decision-support systems. In some cases, as the Stanislav-Petrov incident exemplifies, a healthy level of mistrust of technology can prove valuable. In other cases, a disregard for AI-supported decisions could jeopardize possible benefits of emerging technologies—such as improved situational awareness. Additionally, the immaturity of AI technology could lead to safety and reliability problems (Klare, 2020). There is uncertainty about the predictability and dependability of AI-supported systems due to a lack of transparency with regard to the algorithms

⁷In an interview, Lt. Gen. Jack Shanahan, director of the Joint Artificial Intelligence Center (JAIC), emphasized the US military’s skepticism about integrating AI in NC3 systems: “You will find no stronger proponent of integration of AI capabilities writ large into the Department of Defense, but there is one area where I pause, and it has to do with nuclear command and control” (Freedberg, 2019).

used. Decisions (or suggestions) by AI-supported systems can be opaque and incomprehensible to human operators. Indeed, AI (particularly ML) works like a black box that obscures the way “a system came to a certain conclusion (. . .) Once the operation of AI systems is triggered, humans are unable to monitor the systems’ decision calculus in real time” (Fitzpatrick, 2019). The potential unpredictability of AI-supported systems also makes situational assessments more difficult. This could encourage misjudgments that would be particularly dangerous during nuclear confrontations.

AI-supported systems still lack the flexibility of humans to react to events in broader contexts or to unforeseen incidents (Horowitz et al., 2019; Klare, 2020). While ML could be helpful in some parts of the nuclear enterprise (e.g., detection of anomalies in large data sets), it may prove hard to take advantage of it in many others. Since nuclear weapons have only been used twice—in Hiroshima and Nagasaki in 1945—there is simply not sufficient data to create stable early warning or threat-detection systems, nor data on the performance of these very systems in crises. Ensuing data biases could indeed render the possibility of nuclear confrontation more likely when nuclear risks are overestimated (or underestimated) due to the lack of appropriate data. As a result, “[n]uclear-armed states would have to crack difficult testing issues associated with the design of these systems to be confident that they can be used in a predictable and reliable manner and be certified for use” (Boulanin, 2018). Just as problematic as insufficient data are data biases that are unintentionally introduced into AI-based systems through their learning processes (United Nations Institute for Disarmament Research, 2018). As algorithms and data selection for AI systems and ML are dependent on human definition, pre-existing and implicitly held beliefs may be reflected in both algorithms and training data (Fitzpatrick, 2019)—a case of “you only hear what you want to hear.” The probability of an AI-based early warning system delivering a false positive could increase, for instance, if the system were fed with data reproducing the belief that opponents will indiscriminately carry out nuclear strikes in conflicts without compunction.

Finally, the brittleness and vulnerability of AI-supported and autonomous systems can be critical. The capabilities offered by emerging technologies could be exploited by cyberattacks, jamming, or hacking. Data could be intentionally “poisoned” by skilled (and malevolent) actors to deceive, disrupt or impair NC3 systems (Fitzpatrick, 2019; Klare, 2020). If the exact ways in which machines learn remain black boxes, detecting the exact location of poisoned data also remains difficult. Furthermore, targeted disinformation—such as deep fakes created by capable AI—could be used to either fool human operators or to manipulate decision-making at the height of conflicts, possibly resulting in misjudgments and inadvertent nuclear weapons use (Sauer, 2019). As matters stand now, the greater the dependence on AI, the greater the possibility of critical interference.⁸ Fortunately, the destructive potential of nuclear weapons currently seems to be inducing caution in military

⁸So far, the nuclear powers have tried to reduce these perils: Reportedly, the US nuclear forces are still coordinated via floppy disks to minimise the risks associated with the increasing

professionals and “a strong organizational bias towards maintaining positive human control over nuclear weapons is likely to mitigate against any risks” (Horowitz et al., 2019). With good reason: Inadvertent use of nuclear weapons would (most likely) entail the severest consequences and would (almost certainly) result in the death of millions of people. Thus, it is not surprising that the nuclear field is renowned for its conservativeness and has a record for only slowly integrating new technologies.

2.3 The Nuclear-Conventional Nexus and Nuclear Psychology

Many observers consider the developments and increasing inclusion of AI in conventional weapon systems to be more dangerous in terms of nuclear stability than the use of AI in nuclear weapon systems. The qualitative improvement of conventional weapons through the use of AI or a greater degree of autonomy could jeopardize nuclear second-strike capabilities or the survivability of nuclear forces. Enhanced accuracy could make targets that were previously reserved for nuclear strikes (such as hardened silos for intercontinental ballistic missiles) vulnerable to high-precision conventional weapons (Cuihong, 2019; Verbruggen, 2020). Remote sensing, improved satellite imagery and systematic scanning of ISR data could render mobile missile launchers or sea-based deterrence useless (Cuihong, 2019), for example, by being located and thus potentially exposed to a crippling first strike. Moreover, the thought of having superior capability could lead to riskier strategies. In turn, the fear that an adversary might undermine your own deterrent (or might to do so in the future) could encourage the premature use of nuclear weapons or the equipping of nuclear weapons systems with AI components that have not been properly tested. Furthermore, technological and strategic advances in drone warfare—most importantly swarming—pose a challenge to missile defense systems. In order to compensate for its (at least quantitative) nuclear inferiority, China especially is investing in research on drone swarms that could counter US missile defense in Asia (Saalman, 2019). Not only the higher accuracy (or mass), but also the sheer speed of AI-enhanced or autonomous conventional weapons—both in terms of processing power and physical speed—could increase the risk of nuclear deployment. High-speed weapons such as hypersonic gliders or fully autonomous weapon systems leave little room for deliberate and careful decision-making or de-escalation (Klare, 2020). In order to prevent attacks on their nuclear deterrent, states that doubt the viability of their second-strike capabilities could resort to risky strategies such as putting nuclear weapons on hair-trigger alert (Horowitz et al., 2019). The convergence of nuclear and conventional systems (paired with the increasing speed and accuracy of conventional weapons) poses yet another problem

interconnectedness of their systems (Mangan, 2016) although the US is currently in the process of modernising its information and communications technology (Boulanin et al., 2020, p. 21).

(Klare, 2020). In time-critical situations, a non-nuclear-armed strike from a weapons platform that is capable of launching both nuclear and conventional attacks could potentially be misinterpreted as a nuclear first strike (Gompert & Libicki, 2019). Such scenario may easily result in nuclear retaliation.

The risks associated with the use of AI or autonomy in nuclear weapon systems depend not only on technological capabilities and advances, but also on cognition and the assessment of an adversary's capabilities (Geist & Lohn, 2018). The psychological influence of AI on nuclear strategy is closely linked to the ambiguity of the actual potential of new technologies as well as the lack of transparency about the efforts of nuclear-weapon states to incorporate AI and autonomy into their nuclear weapon systems. Uncertainty about their own technological abilities in comparison with those of others could cast doubt on second-strike capability. Technological pressure (or the "need to catch up") may lead to qualitative arms races and further crisis instability if nascent technology is deployed without testing. Misperceptions about an opponents' capabilities or intentions and resulting miscalculations under conditions of urgency and ambiguity may trigger nuclear escalation (Gompert & Libicki, 2019). Furthermore, implicit assumptions concerning the intentions of other states could be built into nuclear weapon systems through software codes and data input. AI, seemingly a rational and objective *trusted adviser* (Geist & Lohn, 2018), could then fall prey to human biases—and thus reproduce and reinforce pre-existing beliefs that would be reflected in the behavior of AI-based early warning systems or decision-support programs.

In terms of nuclear strategy, the further integration of new technologies into nuclear weapon systems could be particularly attractive to less powerful states (Horowitz et al., 2019). Suspicions concerning the adoption of new AI capabilities by other nuclear-weapon states might lead states to resort to destabilizing measures if they believed that their retaliatory capability were weakened. These destabilizing measures could include further efforts to modernize nuclear arsenals (by including AI or increasing the autonomization of nuclear forces), but also to abandon no-first-use policies, raise the alert statuses of nuclear forces, and further automate nuclear launch policies (Boulanin, 2018; Geist & Lohn, 2018). In addition to the thwarting of nuclear-weapon states' deterrent and second-strike capabilities, this may ultimately weaken the nuclear taboo⁹ and thus render the use of nuclear weapons more likely. Given the cloudiness of the nuclear field and the myths surrounding the capabilities of (AI-)enhanced nuclear weapons, states are dependent on (sometimes imprecise) intelligence and their own assessment of what (other) nuclear-weapon states can do or have (or might have in the future). Yet, "[t]hat is where the inherent nature of AI technology becomes a major problem: the fact that it is software-based makes tangible evaluation of military capabilities difficult. Nuclear-armed states

⁹The nuclear taboo or the tradition of 'non-use' of nuclear weapons refers to the existence of a norm on the prohibition of nuclear weapons use, which has effectively prevented states from using nuclear weapons since WW2 (Tannenwald, 1999).

could therefore easily misperceive their adversaries' capabilities and intentions" (Boulanin, 2019).

3 Opportunities for Nuclear Verification

The use of AI—and its perceived capabilities—could not only threaten but also enhance nuclear stability. It could equally well provide nuclear-weapon states with better information and better decision-making tools for time-critical situations, or it could reduce the risk of miscalculation and accidental escalation. The main prerequisite for the beneficial use of AI is transparency about the actual capabilities of AI-enhanced systems and the intentions with which AI and autonomy are used in nuclear weapon systems. Mutual misperceptions of each other's (technological) potentials and resulting arms spirals could lead to unintended nuclear use. For this reason, political and military leaders should continue to regard AI-enhanced weapon systems with caution, as the premature adoption of insufficiently tested systems could have serious consequences. Another condition for the use of AI in nuclear verification is that it must not make attacks easier, for instance by revealing the existence or location of mobile launchers or a sea-based deterrent.

Importantly, the use of AI also generates new opportunities for the arms control community to monitor nuclear weapon-related developments and conduct verification operations (Kaspersen & King, 2019). A future is imaginable with not only technologically enhanced nuclear weapons but also with improved verification systems that help monitor nuclear nonproliferation, nuclear disarmament or numerical and weapon-related restrictions. Nuclear verification largely relies on the interpretation and evaluation of large sets of data. Hence, AI could contribute to verifying nuclear disarmament, for instance with regard to the detection and characterization of nuclear material, while it could support the exposure of clandestine nuclear proliferation through enhanced data processing and pattern recognition (Exline, 2020; International Atomic Energy Agency-IAEA, 2017; Patton et al., 2016). By supporting pattern recognition or by filtering out unusual movements or developments, AI could help improve the cross-analysis of ISR data, for example to help monitor observance of treaty declarations. Existing and previous nuclear arms control treaties are largely based on telemetry, (satellite) surveillance of declared facilities, and data exchange. Improved satellite imagery could provide better information on troop movements and missile deployment or with regard to weapon-specific, numerical or local restrictions (such as those that were laid down in the 1987 INF (Intermediate-Range Nuclear Forces) Treaty for short-/mid-range and cruise missiles). It could also contribute to monitoring or detecting the transformation of nuclear sites or nuclear weapons silos (Boulanin, 2019; Lück, 2019; Patton et al., 2016). Even now, the US is reported to be developing improved satellite imagery capacities that will enable it to detect mobile (nuclear-capable) missile launchers (Stewart, 2018). Such technology could easily be used for treaty verification. Autonomy, on the other hand, could prove helpful in screening observance of

treaties, especially in inhospitable environments such as the deep sea. In such areas, higher levels of autonomy in monitoring systems (and surveillance vehicles) could enable sensing operations in remote spaces too (Boulanin, 2018). However, the use of AI in verification systems must be carefully balanced. Verification systems should not be designed to potentially make attacks easier, as this would both contradict the very idea of verification as a stabilization of military relations and reduce the political will to contribute to such systems. Here, a principled tension arises between the use of AI for verification of nuclear arms control and for verification of nonproliferation and disarmament. While in the former case real-time location information (e.g., to detect launch platforms) would have to be avoided to prevent competitive advantages for one side and to ensure stability between the parties concerned, this data could improve treaty verification when it comes to nuclear nonproliferation or disarmament.

With regard to the prevention of nuclear proliferation, AI could contribute to the detection of enrichment activities and the search for evidence of the concealment of facilities. The analysis of satellite imagery is already an important part of nuclear nonproliferation endeavors (Niemeyer & Ruthkowski, 2016). AI could be helpful in the analysis of large quantities of satellite imagery: it could be used to measure the intensity of use, energy production and distribution in declared facilities by monitoring temperature or optical differences and to screen for possible modifications of these facilities. Furthermore, improved image recognition and processing through AI could advance the detection and classification of (undeclared) nuclear facilities such as uranium mills (Exline, 2020; Patton et al., 2016). Proliferation threats can also be identified through the analysis of trade data (for example import and export declarations) “to look for indications of technology transfers related to nuclear weapons production, or for transfers involving entities linked to such production” (FAS, 2017). This would be facilitated by AI as it would no longer need to be done manually. The same applies to the tracking and analysis of proliferation networks (Exline, 2020). The IAEA is exploring ways of including AI in its safeguard verification regime (IAEA, 2020). This is most promising in terms of data processing or the use of enhanced satellite imagery. AI and ML could improve existing approaches to material accountancy, the screening of technical data (or the search for anomalies in technical data) or image recognition that is needed to monitor nuclear facilities (Rockwood et al., 2019). However, the use of AI cannot (currently) replace or even eliminate the need for manual work such as the tagging of verified material or instruments.

As impressive as the benefits of using AI and autonomy in nuclear weapon systems can be, the improvement of nuclear verification tools through new technologies is equally appealing. However, the verification of nuclear arms control or nuclear disarmament is and has always been a highly politicized issue. In the past, potential technological advancements have been rejected on political grounds—in cases where the technology was perceived as too intrusive (Lück, 2019). The field of nuclear disarmament verification is particularly delicate in this respect. The verification of nuclear disarmament has to contend with potential knowledge gaps—nuclear-weapon states will most probably not allow algorithms to help verify

disarmament obligations or allow non-nuclear weapon states to participate in disarmament endeavors if these systems are vulnerable to the proliferation of nuclear knowledge. As with the use of AI in nuclear weapon systems, verification systems based on AI technology may be vulnerable to technological intrusion. Hacking, jamming or deep fakes could be employed to disguise treaty violations; (training) data could be deliberately manipulated to restrict the operability of verification systems. The potential vulnerability of AI to these deficiencies makes its use in nuclear weapons verification possible only when supplemented by other verification instruments or human control (IAEA, 2017). Finally, the use of AI can not only improve existing verification tools but can also cause verification problems. AI-enhanced systems or autonomy raise accountability issues: in the future, who will be responsible for nuclear attacks—politicians or autonomous decision-support systems? Questions such as who assumes the liability for nuclear risks, but also of how to verify software codes of AI-enhanced or autonomous nuclear weapons or how to ensure a minimum level of stability and safety, are not unique to the use of AI in nuclear weapons. However, these problems are exacerbated by the sheer destructiveness of such weapons.

4 Advanced Wonder Weapon or Doomsday Machine?

The inclusion of AI and autonomy in nuclear weapon systems can indeed be a two-edged sword. New technology is not dangerous in itself. However, in the face of recurrent nuclear confrontation and the looming threat of a multilateral nuclear arms races, the use of AI and autonomy in nuclear weapon systems could prove to be the opening mechanism for a (nuclear) Pandora's box. Faster and more complex decision cycles, shorter reaction times in crises, and new first-strike advantages could shake up traditional strategic planning of nuclear warfare and military operations. In such scenarios, governments and military could underestimate or disregard the limitations of current technology when confronted with a perceived strategic imbalance. The hasty inclusion of AI technology in nuclear weapon systems could aggravate already existing risks and provoke a qualitative as well as a quantitative nuclear arms race. However, a technological arms race in the nuclear realm could be hazardous: "racing blindly down the path toward 'smarter' weapons, with nuclear risks remaining as inadequately addressed as they are now, might well turn the military applications of AI and machine learning into a shortcut to Armageddon" (Sauer, 2019). Then again, better satellite technology, image processing and the possibility of examining even larger sets of data could make nuclear weapon systems more reliable, increase crisis awareness and contribute to the development of new and more dependable instruments for nuclear arms control and verification.

It is important for us to understand better what possibilities exist. Discussions of the benefits and perils of nuclear applications of AI, the (qualitative) modernization of nuclear weapons arsenals, and the impact of technological advances on nuclear strategy are still at a fledgling stage. In order to deepen this discussion, it is necessary

to explore which kinds of technology states are developing, what they already have and what they are planning. This discussion should not be restricted to academic circles; it has to include a dialogue among those states that are increasing the use of AI in nuclear weapon systems: “States need to not only develop and [gain a] better understanding [of] the opportunities and challenges posed by the military use of AI, particularly in the nuclear force-related context; they also need to discuss these with other states” (Boulanin, 2019). Furthermore, tailored arms control endeavors are necessary in order to reduce the risks associated with the use of new technologies in nuclear weapon systems. This could include both confidence-building measures and dialogue on the technical security of the use of AI or autonomy in nuclear weapon systems to mitigate against accidental use, false negatives (or positives) or miscalculations. Transparency on the actual capabilities of emerging technologies in the nuclear field could reduce some of the risks associated with AI, since assessing an opponents’ capabilities is a highly psychological act. This means that perceptions of what AI-enhanced weapon systems could do can have just as many concrete effects as the actual capability of those very weapons—even (or especially) if what is imagined reaches far beyond reality. Transparency in military planning with regard to nuclear applications of AI is only one step; avoiding the entanglement of nuclear and conventional systems is another. The more conventional weapon systems are enhanced with both AI and possibly nuclear firepower, the greater the risk of inadvertent nuclear escalation. Generally, appropriate risk reduction measures in the nuclear field do not change just because AI is involved. More transparency on nuclear doctrines, more defensive nuclear doctrines, lower alert status, and no-first-use policies are essential for keeping the risk of nuclear use as low as possible. If anything, these measures are all the more important the more nuclear-weapon states rely on AI-based systems, as any weakening of the nuclear taboo could have devastating consequences.

Finally, we need more research on the prospects of AI in nuclear disarmament, arms control, and nonproliferation verification. Currently, most research in the nuclear field focuses on the benefits and dangers of including AI in nuclear weapon systems and not on its more “friendly” applications. While the IAEA announced that it would carry out research on how to incorporate AI and ML into its safeguard verification regime, this work seems to be progressing only slowly. Nevertheless, this very area would be a good opportunity for non-nuclear-weapon states to become involved and not leave discourse about the nuclear field only to nuclear-weapon states.

References

- Atherton, K. (2020, January 3). *Will Russia’s nuclear-armed bombers in 2040 be drones?* C4ISRNET. Retrieved from <https://www.c4isrnet.com/unmanned/2020/01/03/will-russias-nuclear-armed-bombers-in-2040-be-drones/>

- Borrie, J. (2019). Cold war lessons for automation in nuclear weapon systems. In V. Boulanin (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk, Euro-Atlantic perspectives* (pp. 41–52). Stockholm International Peace Research Institute (SIPRI).
- Boulanin, V. (2018, December 7). *AI and nuclear weapons—Promise and perils of AI for nuclear stability*. UNU-CPR Centre for Policy Research. Retrieved from <https://cpr.unu.edu/ai-global-governance-ai-and-nuclear-weapons-promise-and-perils-of-ai-for-nuclear-stability.html>
- Boulanin, V. (2019). *The impact of artificial intelligence on strategic stability and nuclear risk*. Stockholm International Peace Research Institute.
- Boulanin, V., Saalman, L., Topychkanov, P., Su, F., & Peldán-Carlsson, M. (2020). *Artificial intelligence, strategic stability and nuclear risk*. Report. Stockholm International Peace Research Institute. Retrieved from <https://www.sipri.org/publications/2020/other-publications/artificial-intelligence-strategic-stability-and-nuclear-risk>
- Boulanin, V., & Verbruggen, M. (2017). *Mapping the development of autonomy in weapon systems*. Report. Stockholm International Peace Research Institute. Retrieved from <https://www.sipri.org/publications/2017/other-publications/mapping-development-autonomy-weapon-systems>
- Cuihong, C. (2019). The shaping of strategic stability by artificial intelligence. In L. Saalman (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk, East Asian perspectives* (pp. 54–77). Retrieved from <https://www.sipri.org/publications/2019/other-publications/impact-artificial-intelligence-strategic-stability-and-nuclear-risk-volume-ii-east-asian>
- Exline, P. (2020). Machine learning in the countering weapons of mass destruction fight. In M. Kosal (Ed.), *Disruptive and game changing technologies in modern warfare* (pp. 71–92). Springer.
- Federation of American Scientists. (2017, September). *Nuclear monitoring and verification in the digital age: Seven recommendations for improving the process*. Third report.
- Fitzpatrick, M. (2019). Artificial intelligence and nuclear command and control. *Survival*, 61(3), 81–92.
- Freedberg, S. (2019, September 25). No AI for nuclear command & control: JAIC's Shanahan. *Breaking Defense*. Retrieved May 27, 2020, from <https://breakingdefense.com/2019/09/no-ai-for-nuclear-command-control-jaics-shanahan/>
- Gady, F. S. (2019a, February 20). Russia's first 'Poseidon' underwater drone-carrying submarine to be launched in 2019. *The Diplomat*. Retrieved from <https://thediplomat.com/2019/02/russias-first-poseidon-underwater-drone-carrying-submarine-to-be-launched-in-2019/>
- Gady, F. S. (2019b, March 26). US intelligence: Russia's nuclear-capable 'Poseidon' underwater drone ready for service by 2027. *The Diplomat*. <https://thediplomat.com/2019/03/us-intelligence-russias-nuclear-capable-poseidon-underwater-drone-ready-for-service-by-2027/>
- Geist, E., & Lohn, A. J. (2018). *How might artificial intelligence affect the risk of nuclear war?* Publication. RAND Center for Global Risk and Security.
- Gertler, J. (2019). *Air force B-21 raider long-range strike bomber*. Report No. R44463. Congressional Research Service.
- Gompert, D., & Libicki, M. (2019). Cyber war and nuclear peace. *Survival*, 61(4), 45–62.
- Horowitz, M. C. (2019). Artificial intelligence and nuclear stability. In V. Boulanin (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk, Euro-Atlantic perspectives* (pp. 79–83). Retrieved from <https://www.sipri.org/publications/2019/other-publications/impact-artificial-intelligence-strategic-stability-and-nuclear-risk-volume-i-euro-atlantic>
- Horowitz, M. C., Scharre, P., & Velez-Green, A. (2019). *A stable nuclear future? The impact of autonomous systems and artificial intelligence*. Working Paper, arXiv: 1912.05291. <https://arxiv.org/abs/1912.05291>
- International Atomic Energy Agency. (2017, February). *Emerging technologies workshop*. Vienna.
- International Atomic Energy Agency. (2020, January). *Emerging technologies workshop*. Vienna.
- Kaspersen, A., & King, C. (2019). Mitigating the challenges of nuclear risk while ensuring the benefits of technology. In V. Boulanin (Ed.), *The impact of artificial intelligence on strategic*

- stability and nuclear risk, Euro-Atlantic perspectives* (pp. 79–83). Retrieved from <https://www.sipri.org/publications/2019/other-publications/impact-artificial-intelligence-strategic-stability-and-nuclear-risk-volume-i-euro-atlantic>
- Klare, M. (2020, April). ‘Skynet’ revisited: The dangerous allure of nuclear command automation. *Arms Control Today*. Retrieved from <https://www.armscontrol.org/act/2020-04/features/skynet-revisited-dangerous-allure-nuclear-command-automation>
- Lück, N. (2019). *Machine learning-powered artificial intelligence in arms control*. PRIF Report 8/2019. Peace Research Institute Frankfurt.
- Mangan, D. (2016, May 26). US military uses 8-inch floppy disks to coordinate nuclear force operations. *CNBC*. Retrieved from <https://www.cbc.com/2016/05/25/us-military-uses-8-inch-floppy-disks-to-coordinate-nuclear-force-operations.html>
- Niemeyer, I., & Ruthkowski, J. (2016, June). *Satellite imagery processing for the verification of nuclear non-proliferation and arms control*. In Paper presented at European Association of Remote Sensing Laboratories Symposium Bonn.
- Patton, T., Lewis, J., Hanham, M., Dill, C., & Vaccaro, L. (2016). *Emerging satellites for non-proliferation and disarmament verification*. Vienna Center for Disarmament and Non-Proliferation.
- Rockwood, L., Mayhew, N., Lazarev, A., & Pfnaisl, M. (2019) *IAEA safeguards: Staying ahead of the game*. Swedish Radiation Safety Authority Report 14/19. Vienna: Vienna Center for Disarmament and Non-Proliferation.
- Saalman, L. (Ed.). (2019). *The impact of artificial intelligence on strategic stability and nuclear risk: East Asian perspectives*. Stockholm International Peace Research Institute.
- Sauer, F. (2019). Military applications of artificial intelligence: Nuclear risk redux. In V. Boulanin (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk* (pp. 84–90). Stockholm International Peace Research Institute.
- Sayler, K. (2019). *Artificial intelligence and national security*. Report No. R45178. Congressional Research Service.
- Scharre, P., & Horowitz, M. (2018). *Artificial intelligence. What every policymaker needs to know*. Center for a New American Security. Retrieved from <https://www.cnas.org/publications/reports/artificial-intelligence-what-every-policymaker-needs-to-know>
- Schörnig, N. (2018). *Preserve past achievements. Why drones should stay within the missile technology control regime*. Report No. 149. Peace Research Institute Frankfurt.
- Stefanovich, D. (2020). Artificial intelligence advances in Russian strategic weapons. In P. Topychkanov (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk: South Asian perspectives* (pp. 25–36). Stockholm International Peace Research Institute.
- Stewart, P. (2018, June 5). Deep in the pentagon, a secret AI program to find hidden nuclear missiles. *Reuters*. Retrieved from <https://www.reuters.com/article/us-usa-pentagon-missiles-ai-insight-idUSKCN1J114J>
- Tannenwald, N. (1999). The nuclear taboo. The United States and the normative basis of nuclear non-use. *International Organization*, 53(3), 433–468.
- Topychkanov, P. (2019). Autonomy in Russian nuclear forces. In V. Boulanin (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk* (pp. 68–76). Stockholm International Peace Research Institute.
- United Nations Institute for Disarmament Research. (2018). *Algorithmic bias and the weaponization of increasingly autonomous technologies*. UNIDIR Resources No. 9.
- Verbruggen, M. (2020). The extensive role of artificial intelligence in military transformation. In P. Topychkanov (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk: South Asian perspectives* (pp. 11–16). Stockholm International Peace Research Institute.

AI, WMD and Arms Control: The Case of Nuclear Testing



Anna Heise

● آقای هوش مصنوعی ●

رسانه هوش مصنوعی دانشگاه تهران

@MrArtificialintelligence

Abstract Although actual nuclear tests have decreased considerably in recent decades, this weapon technology still exerts a great attraction. However, advances in the simulation of nuclear detonations, the increased use of AI to predict their effects, and AI-assisted improvements in the analysis of the vast amounts of nuclear test data are threatening this trend. At the same time, however, AI also offers the possibility of significantly improving existing ways of deducing tests based on seismic information or the measurement of radionuclides, and of making predictions or estimates based on incomplete data. The chapter addresses these developments and challenges and analyses the technical as well as organizational and structural changes for nuclear weapons testing, but also their containment, control and monitoring that could result.

1 The Nuclear Testing Status Quo

Over the past decades, the number of nuclear tests performed by various states has plummeted from several events per week (1980s) to approximately one per year (2006–2018, North Korea) to none at present. This was achieved mostly by multi-lateral treaties which initially banned nuclear explosions in the atmosphere, outer space and under water—the Partial Test Ban Treaty (PTBT) (Treaty Banning Nuclear Weapon Tests in the Atmosphere, in Outer Space and Under Water, 1963)—and later prohibited testing completely—the Comprehensive Test Ban Treaty (CTBT) (Comprehensive Nuclear Test-Ban Treaty, 1996). Still, nuclear testing remains a threat, firstly to the environment and secondly as a tremendous proliferation risk because the testing of a nuclear device is crucial for its possible use. In this paper, we will evaluate possible future developments in nuclear testing in our technical era with regard to the use of artificial intelligence (AI). We will discuss changes in the data processing of nuclear tests and its verification and the question of whether this new method for simulating tests might increase the risk of proliferation.

A. Heise (✉)
Hamburg, Germany

2 Application of AI in Nuclear Testing

AI—especially in the context of lethal autonomous weapon systems—is sometimes referred to as the third revolution in military technology, after gunpowder and nuclear weapons (Russell et al., 2015). While its application varies across a wide spectrum from language processing to deep learning and robotics, most important in the context of nuclear tests is machine learning (ML). For nuclear testing, all three major types of ML—regression, classification and clustering—can be applied in data analysis. Monitoring stations that are operated under the CTBT gather a huge amount of information. This data needs to be sorted, reduced, and categorized so signals from nuclear and from other events can eventually be distinguished. Another possible application of AI for nuclear testing is robotics, where a possible use case could be the collection of field samples of possibly contaminated air or ground samples, for instance by AI-enabled drones or small vehicles. The vehicles could either be used for reconnaissance and verification operations after the testing of a nuclear device or by the testing party itself, since it allows data gathering in what are probably highly inaccessible areas at the test site. These different aspects—with a focus on ML—and their different applications will be discussed in detail in the following sections.

2.1 *AI and Virtual Testing*

The first nuclear weapon detonation took place in the race for military dominance and was based on experimental engineering rather than code-based simulations. Almost 20 years later, following the PTBT, nuclear tests by the signature states went underground in 1963. In the United States (US), nuclear tests finally halted indefinitely in 1992 and for scientists working at the National Laboratories the CTBT ended a way of life. McNamara (2001, p. 114) writes: “The design and test cycle acted as an engine for the ongoing integration of expertise and the social reproduction of the weapons community; indeed, experimental activity was critical in organizing social relations among the hundreds of staff members involved in weapons work at Los Alamos.” This highlights the importance of the human factor, which has been at center stage in earlier phases of developing and maintaining nuclear weapons. The abandoned testing regime was followed by the Stockpile Stewardship Program (SSP), with its main focus on maintaining the US nuclear arsenal (Gordin et al., 2005) without actually detonating a device. As part of the SSP, various facilities at different sites carry out different tasks. For instance, the Dual-Axis Radiographic Hydrotest Facility (DARHT) performs subcritical tests by using x-rays to understand the implosion process in a nuclear bomb in three dimensions. ATLAS (Advanced Tracking Laser Alignment System) at the US National Ignition Facility (NIF) has the objective of inducing nuclear fission by compression of hydrogen fuel with lasers. The Accelerated Strategic Computing Initiative (ASCI)

is a program for replacing old supercomputers with parallel computing systems to change from two- to three-dimensional codes (Gordin et al., 2005). The transition from experimental testing to simulation began in the early 1990s and lasted until the turn of the century.

This change from real-life explosions to computer simulations brought new challenges for the programmers: Suppose that the explosion yielded several percent less power than predicted by the model. If the developers understood the underlying mistake they would of course adapt the physical model, but if this was not the case they inserted a factor (called a *fudge* factor or knob) to bring the model into agreement with the experimental data (Gordin et al., 2005). Sometimes these code adjustments are the product of the developers' intuition and not based on understanding the underlying physics. This makes the theoretical, computer-based approach complicated, since it is still highly reliant on the human expert. These circumstances led to a different kind of nuclear pedagogy in the training of future nuclear engineers today (Gordin et al., 2005). The procedure is similar to the learning processes of AI, which indicates that the future engineer might as well be replaced with a ML algorithm. According to Baker (1997): "As a nuclear weapons designer I learned the limitations of simulations and the humility that comes with the failure of a nuclear test."

Computer calculations, regardless of how good or fast the computer is, are only as good as the data and models you give them and the knowledge and experience of the person doing the calculations. Even today no computers are big enough or fast enough to simulate all that goes on when a nuclear weapon explodes. True understanding of and experience with the limitations of calculations came from understanding the differences between calculations and experiments, including nuclear tests. The general shift from human operator-based experimental tests to computer-driven virtual testing means not only a change in method but also in a shift in the work culture and, as a result, is rejected by some (especially older-generation nuclear weapons designers). Future developments in the use of AI in the context of simulated nuclear testing will thus not only be dependent on the technology but on the attitude of those in charge of change. One of the main reasons for extensive nuclear testing in the past was the immediate impression this demonstration of military power made on opponent states. Since virtual testing is kept secret, its only purpose is to serve scientific and engineering interests. The data gathered in test explosions is highly confidential and any leakage of knowledge is considered a major threat.

With this shift from experimental to virtual testing, which includes different AI mechanisms, it might be asked whether virtual testing based on AI is easier. The cost of nuclear weapons research and development in the US is higher at present than it was during the Cold War (when the stockpile was larger and nuclear tests more frequent). A major part of the budget is spent on expensive simulation technology, which is used to simulate components of nuclear weapons. Another aspect is less scientific but might be the most important: the fact that simulating a nuclear explosion is an abstract endeavor. Researchers reported a considerable loss of excitement and thrill when nuclear tests stopped being conducted above ground and were taken underground. This connection between event and action is even

further weakened by a theoretical simulation without any actual explosion. Successes with virtual testing based on AI will further increase this gap as well as the already existing absence of any ethical considerations or emotional connection. This might heighten the efficiency of testing itself at the expenses of a greater opportunity for making the actual use of a nuclear weapon politically acceptable (the so-called *dulling effect*).

Further Reading Since discussion of the topic of virtual testing is restricted, only a few publications are available. We suggest Gordin et al. (2005) who give a comprehensive overview of the changing methods over time and the role of the engineer. It is well suited for experts outside physics or engineering.

3 Detecting Nuclear Tests: Present State of the Art

Since 1996 the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) installed the first suite of sensors in the international monitoring system (IMS), the network has expanded (currently 337 stations worldwide), and the technology has been developed further. Monitoring stations have been updated as well as event diagnostics enhanced. The IMS is a global network of different monitoring stations. It includes waveform physical sensor stations (seismic, hydroacoustic and infrasound) as well as radionuclide stations, of which half are equipped with noble gas-detection equipment. The latter is most important for verification purposes, since the nuclear character of an event can only be confirmed by detecting radioactive noble gas isotopes, usually xenon and argon. The monitoring stations are connected via a global network to a centralized processing system at the International Data Centre (IDC) in Vienna. The IDC operates continuously and in real time. Two aspects are relevant: The processing of single-station data, in which the raw seismic sensor data is reduced and analyzed to detect and classify incoming signals at the corresponding station, and network processing, through which signals from the same event are associated with signals collected at different stations. This is done by an automated algorithm and, in a second step, by a data analyst who carries out post-processing to screen the data for additional events or false alarms. Fully automated data processing is currently not possible due to errors caused by the automated method: false detections, detections not occurring due to station noise, incorrect classification of incoming signals, and incorrect associations (Russell et al., 2010). CTBTO data is used not only for verification purposes, but also for investigating the impact of climate change, issuing tsunami warnings or tracking radiation on a global scale. To distribute the raw data the CTBTO has installed the virtual Data Exploitation Centre (vDEC). Scientists can apply to use data desired for a designated project and present their results.

3.1 *Detecting a Nuclear Test with Seismic Waveform Analysis*

The detonation of any device underground creates seismic waves in the ground as the detonation shakes up the soil surrounding the test site. These waves can be detected over long distances with seismic arrays around the globe because their physical nature allows them to travel over long distances. This is even more true for hydroacoustic waves since water transmits acoustic waves even better. The signals caused by an explosion differ from those of naturally occurring events such as earthquakes and allow a precise localization of the event in question to within a few kilometers. Unfortunately, the signal contains no further information making it possible to distinguish a chemical from a nuclear explosion, apart from the actual yield, that can give an indication of what type of event it was. To verify the nuclear character of an event, traces of radioactive isotopes need to be measured. This can be done by means of different procedures.

3.1.1 **Creating a High-Quality Event Bulletin**

The IMS consists of stations for different waveform technologies which are seismic, hydroacoustic and infrasound. Since the most recent nuclear tests have all been performed underground, seismic stations are of high interest. These stations operate constantly and collect data which is sent to the IDC for further analysis and interpretation. The aim is to create an event bulletin which logs as a possible detection any event that occurs and can be used for any further interpretation and (eventually) decision-making.

At the IDC the data that has been collected is processed in the following way: The data is filtered several times by the automatic detection system and the resulting events are presented in a list (named Standard Event List 3—SEL3). This list is then reviewed by analysts to eliminate any false positives (events which the system labeled as detections although they were not). Events that are not screened out by the analysts are known as *good automatic origins* and are labeled *true positives*. Events, that are not on the list but should have been detected by the system are added manually by the analysts and are known as *extra analyst-built origins* and labeled *false negatives* (missed detections). Naturally, this process applies only to events which are represented in the data. In reality there might be more events which were not detected at any stage of the verification process. Such events are regarded henceforth as correct non-detections and are labeled *true negatives* (but can never be measured). The result is a high-quality reviewed event bulletin (REB). The interesting step here is the amount of human effort involved from automated list to event bulletin. Regardless of the number of operating seismic stations, the percentage of automatic events which survive the analysts' review is about 47%. This accounts for about 87% of the final data. The remaining 13% are events missed by the automated processing and are added manually to the bulletin by the analyst.

Procopio et al. (2009) used a ML algorithm with the aim of reducing the output of false events by the system to reduce the workload of analysts. This translates into a two-class problem, where the classifier predicts one of two possible labels for a given event. The supervised learning models were trained on half of the data from SEL3 and then further tested on the remaining data. A decision-tree model was used in the corresponding study to split on the basis of a single attribute of the event. This represents a typical application of a ML algorithm for data reduction. The sheer load cannot be handled by a human analyst and, since screening for events is a purely objective selection process, it can be learned by an algorithm. This method is applied in several different fields of science and economics. Seismic signals are only one example of big data which can be handled efficiently by AI only.

In the future this application will improve even further since the data is numeric and does not include any pattern or complex object which might be difficult to analyze for AI (such as handwriting or faces). When the event bulletin can be created by an ML algorithm without any further improvement by a human analyst, this will be an important step in efficiency, because it will result in greater human capacity for more creative work.

The seismo-acoustic motion created by a near-surface explosion can be used to estimate the yield of the device involved. This approach is complicated by the fact that for proper yield estimation the depth of burial must be known; in most cases the level of uncertainty about this is substantial at best and has a significant impact on the resulting values.

Further Reading Russell et al. (2010) gives a very well-written and comprehensive overview of the different steps in the data processing of seismic signals. Different ML concepts are discussed and explained in an understandable way.

3.1.2 Yield Estimation

An example of the use of AI to extract information from the vast amount of existing data from past nuclear tests is ML algorithms which learn to identify visual features from videos of explosions for further yield-estimation analysis. The nuclear explosion yield of an atmospheric test can be estimated based on 3D visual information. Volume-based models have been demonstrated to reduce the error significantly compared with radius-based models. However, no 3D representations of past tests exist, and new data is not available since atmospheric tests are now banned. Consequently, research focuses on building 3D representations from existing data recorded in the early stages of atmospheric nuclear testing. The US conducted and filmed over 200 atmospheric tests during the 1950s and 1960s. The Lawrence Livermore National Lab digitalized these films which can now be used as validation for fundamental nuclear explosion models. The method currently used for feature detection is scale-invariant feature transform (SIFT). It has become the standard for 3D reconstruction from 2D representations (photographs as well as films), but only works well for continuous angular coverage of the given data. This is problematic

since the available videos of nuclear tests are separated by up to 80°. To merge different videos, hotspots consisting of bomb casing debris are used. This debris is spread across a wide area and varies in size and orientation but may be useful in correlating multiple frames from different camera perspectives. Among other things, Schmitt and Peterson (2014) used a supervised and an unsupervised learning algorithm to find the best feature for detecting hotspots. This has the huge advantage that no further human intervention is needed. Since the algorithm is unsupervised it improves itself on its own and adapts to each new problem that arises.

The results show that the average runtime per image was 1800 times longer for the supervised learning feature. The unsupervised ML algorithm achieved a good overall hit and precision rate and a fairly low false-alarm rate. These results could also be used for other features of interest in explosion data that have been collected and could be applied to different datasets. The detection of hotspots is of interest to different user groups ranging from verification purposes (yield estimation of a particular event) at one end of the spectrum to virtual testing (model evaluation for blast propagation, etc.) at the opposite end.

Whereas ML algorithms have proved to be a valuable tool because of their above-mentioned advantages for analyzing large amounts of data, in a nuclear testing context the bottleneck is the availability of datasets for analysis and training. Because actual nuclear explosions are rare, datasets from earlier events or simulations are the main source for analysis and training.

3.2 *Radionuclides*

The only technology able to distinguish low-yield clandestine nuclear explosions from other events is the monitoring of radionuclides and radioactive noble gases in particular. Currently, four xenon isotopes are being monitored by 40 stations of the IMS around the world. The challenge is to discriminate signatures of a supposed nuclear explosion from (regular) emissions from civilian sources such as nuclear power plants or (especially challenging) medical isotope production facilities. Further difficulties arise from the very limited set of data from past tests: Until now it has never been possible to measure all four isotopes after an actual test. Nevertheless, these datasets are helpful for future comparisons. The established method (Kalinowski et al., 2010) of distinguishing emissions from a potential explosion from other emissions is a log-log graph of the four isotopes sorted into two ratios. The discriminator between the two types of emission in this graph is a line which works independently of the corresponding decay time. This procedure is complicated when new signals with high signal strength are added to the dataset. The linear discriminator breaks down since data which correspond to the non-explosion emission is now located in the explosion section of the plot. As a result, different sorting methods are required.

The task at hand resembles a classification problem, in which the measured concentration can be classified either as an explosion or as background. It is

challenging to discriminate signals in a higher xenon environment, because background and explosion will not be separate measurements but mixed together. Since traditional algorithms which discriminate between all data classes to build models are known to suffer from such an imbalance in the dataset, one-class classifiers are important. In this method, only the data for one class, the target class, is used to build a model, so that this approach relies on recognition rather than on discrimination. In this case, the algorithm is trained on the background dataset, which is highly complex, and modeling the various nuances is difficult. A traditional approach is density estimation in which a statistical distribution is fitted to the target class. The learned density function can then be used to classify values: Those with high density belong to the target class and values of lower density are sorted into the outlier class. Nonetheless, of course, there is a variety of different algorithms, some of which have specifically been developed for one specific class classification, and each has its own challenges and limitations. Sharma et al. (2012) suggested a method in which the data is clustered by an algorithm (k-means) before a one-class classifier is used to model the clustered data. The training dataset for CTBT purposes consisted of a series of simulated industrial radioxenon emitters and random clandestine tests (since there is no real data available for the latter). In total, three CTBT datasets have been used with different ratios of imbalance. All one-class classifiers explored performed better if the data had been clustered first, without much difference in the imbalance between target (background) and outlier (explosion) class.

A more comprehensive approach to ML and xenon measurements was chosen by Stocki et al. (2010). They used a background dataset and a background plus synthetic explosion dataset. The real data was taken from the Ottawa monitoring station since it has one of the highest background rates and hence is one of the most challenging setups for classification. They then reviewed several different ML techniques by building linear and non-linear classifiers. The selected algorithms were: Naive Bayes, Multiple Layer Perceptron, Support Vector Machine, k-Nearest Neighbors and Decision Tree. The latter method is able to explain decisions that are taken, which can be important for discovering new features. To train and test the corresponding algorithm, a statistical method called tenfold cross-validation was used. It was developed for problems with small datasets: The data is stratified into tenfolds and in each run ninefolds are used to train the algorithm and the remaining one for testing. Stocki et al. (2010) found that all ML algorithms outperformed the traditional approach. In their study, the linear discriminator achieved an accuracy level of 57%, whereas the ML algorithms achieved 83% and 98.9% accuracy.

ML is also used to establish new verification methods for the measurement of radioactive argon isotopes. Unlike xenon, argon is not produced in the fission of uranium or plutonium but by activation of stable argon isotopes in the air or of calcium in the soil surrounding the test site [see for example Heise (2019) for a detailed discussion]. Mace and Ward (2018) used a deep-learning neural network to analyze radio-argon signals in an underground laboratory for future verification purposes. Deep learning is one method in the broader family of ML algorithms. It uses a series of layers which perform different tasks. These resemble the activity of

neurons in a biological brain where each layer performs a transformation from input to output. The network was trained on several tens of thousands of signals to distinguish between the detection of radioactive decay and noise. The hit rate has reliably been at about 99%.

Further Reading A detailed discussion of the above-mentioned different ML methods in the context of radionuclide measurement for verification purposes is to be found in Stocki et al. (2010), who discuss the difficulty of discriminating between emissions from tests and other sources as well as the use of different ML algorithms.

3.3 *Threads for AI in Nuclear Test Detection*

Some scientists have mentioned the idea that if the CTBTO collects, processes and releases data of such high quality and quantity, testing a nuclear device itself might become obsolete. However, it must be pointed out that the CTBTO data is used for verification purposes, and it does not allow for any extraction of information about the actual weapon design. No further proliferation risk arises from the CTBTO datasets in combination with ML algorithms. Yet datasets about an actual explosion are of crucial interest to parties which want to develop a nuclear device of their own. A comprehensive collection of data with efficient AI methods might be enough to provide the desired knowledge about the design of a nuclear weapon and in the technical age it might be easier to access the desired data than the corresponding knowledge (whether from a person or a blueprint). Murphy (2019) argues that ML is a current threat in the context of cybersecurity and that the remote measuring stations of the CTBTO are exposed to cyberattacks. The source cited by Murphy (2019) cannot be traced, but the general problem remains that the learning mechanism of AI algorithms relies heavily on the accuracy of the specific events in question. Any changes made by a cyberattack (for example, inserting false positives into the data set) could invalidate the verification mechanism: If the algorithm raises an alarm based on high values which are not true measurements, the station is no longer functioning reliably. Should the dataset be compromised by a malicious attack from a third party, e.g., by tampering with the training dataset, then subsequent conclusions made by the AI would be negatively affected by it and, depending on the attack, might even be completely false. Since real events are so rare, we cannot rely on these events alone to *outbalance* the falsified events. Assuming this attack was performed subtly enough, and not directly noticed, then a series of incorrect conclusions drawn by the AI on the basis of this dataset might be the result. Even worse, other algorithms developed to classify data might be fundamentally wrong if, as a bandwagon effect, their developers tried to bring their results into agreement with these published incorrect methods.

False attribution via a cyberattack will lead to permanent false conclusions, since the false data is integrated into the data pool used for every future decision. Consequently, it is necessary for experienced scientists to check and evaluate the data pool regularly, to ensure that no such false data has been inserted. Given the

complexity of this task, different measures should be used for the evaluation. Finally, the question arises whether analysis will at some point become almost impossibly difficult for humans.

4 Conclusion

All fields in the realm of nuclear testing require the analysis of huge amounts of data and hence depend on the use of ML algorithms to reduce the workload of human analysts and improve the overall quality of the output.

Surprisingly enough, apparently the available technology is being implemented only reluctantly (or the technology has been implemented but the process or results have not been documented or published, for whatever reason). While this is to be expected in the field of virtual testing given its restricted data policies, it cannot apply to the verification of nuclear testing by the CTBTO. It should be noted that several of the available publications concerning the verification of the CTBT which have been used in this work were published about a decade ago. Since then, nuclear testing, as well as technological knowledge and use of AI, have continued to develop during this time. Because of the lack of published material, conclusions drawn here must be taken with a grain of salt. This might be due to the sometimes rigid working structure of the UN as an organization: The system it has in place is hardly ever altered or changed, and if so only quite slowly.

While this might be the case for the CTBTO it certainly is not for national laboratories or other government institutions which deal with the task of providing their government with a working nuclear device. Russell et al. (2010) notes the availability of raw data for testing purposes as the principal obstacle to the implementation of ML. At least for the CTBTO, this will remain an issue which might still be resolved by the use of more efficient training algorithms.

The use of ML and AI, in general, is a major shift in working culture. This has been especially highlighted in the case of virtual testing. While the sheer workload of data analysts is reduced, it throws the scientist back on personal creative abilities and might be a factor in resistance to change. From a technical perspective, ML algorithms are capable of supporting and improving the complex task of data analysis; so far this has mainly been in terms of the amount of data that can be processed, but in the near future the quality of the corresponding output will take on importance as well.

References

- Baker, R. (1997). *Prepared testimony before senate governmental affairs committee, subcommittee on international security*.
- Comprehensive Nuclear Test-Ban Treaty. (1996). UN. <https://www.ctbto.org/verification-regime/background/overview-of-the-verification-regime/>

- Gordin, M., Gooday, G., Gusterson, H., & Ito, K. (2005). *Pedagogy and the practice of science: Historical and contemporary perspectives*. MIT University Press.
- Heise, A. C. (2019). *Machbarkeitsstudie zur Nutzung des Radioisotopes Argon-37 im Rahmen des Verifikationsregimes des Umfassenden Kernwaffenteststopp-Verrtrags*. Universität Hamburg.
- Kalinowski, M. B., Axelsson, A., Bean, M., Blanchard, X., Bowyer, T. W., Brachet, G., Hebel, S., McIntyre, J. I., Peters, J., Pistner, C., Raith, M., Ringbom, A., Saey, P. R. J., Schlosser, C., Stocki, T. J., Taffary, T., & Kurt Ungar, R. (2010). Discrimination of nuclear explosions against civilian sources based on atmospheric xenon isotopic activity ratios. *Pure and Applied Geophysics*, 167(4–5), 517–539. <https://doi.org/10.1007/s00024-009-0032-1>
- Mace, E., & Ward, J. (2018). *Enhanced detection of nuclear events, thanks to deep learning*.
- McNamara, L. A. (2001). *Ways of knowing about weapons: The cold war's end at the Los Alamos National Laboratory*. University of New Mexico Albuquerque.
- Murphy, M. (2019). *The cybersecurity protection of peacetime organizations: Comprehensive test ban treaty organization*. Utica College.
- Procopio, M., Young, C. J., & Lewis, J. E. (2009). *Using machine learning to improve the efficiency and effectiveness of automatic nuclear explosion monitoring systems*. National Nuclear Security Administration.
- Russell, S., Hauert, S., Altman, R., & Veloso, M. (2015). Robotics: Ethics of artificial intelligence. *Nature*, 521(7553), 415–418. <https://doi.org/10.1038/521415a>
- Russell, S., Vaidya, S., & Le Bras, R. (2010). *Machine learning for comprehensive nuclear-test-ban treaty monitoring*. CTBTO Spectrum. https://www.academia.edu/1148521/Machine_learning_for_comprehensive_nuclear_test_ban_treaty_monitoring
- Schmitt, D. T., & Peterson, G. L. (2014). *Machine learning nuclear detonation features*. In 2014 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) (pp. 1–7). <https://doi.org/10.1109/AIPR.2014.7041936>
- Sharma, S., Bellinger, C., & Japkowicz, N. (2012). Clustering based one-class classification for compliance verification of the Comprehensive Nuclear-Test-Ban Treaty. In L. Kosseim & D. Inkpen (Eds.), *Advances in artificial intelligence. Canadian AI 2012. lecture notes in computer science*, 7310 (pp. 181–193). Springer. https://doi.org/10.1007/978-3-642-30353-1_16
- Stocki, T. J., Li, G., Japkowicz, N., & Ungar, R. K. (2010). Machine learning for radioxenon event classification for the Comprehensive Nuclear-Test-Ban Treaty. *Journal of Environmental Radioactivity*, 101(1), 68–74. <https://doi.org/10.1016/j.jenvrad.2009.08.015>
- Treaty banning nuclear weapon tests in the atmosphere, in outer space and under water*. (1963). U.N.T.S. <https://treaties.un.org/pages/showDetails.aspx?objid=08000002801313d9>

Artificial Intelligence in Conventional Arms Control and Military Confidence-Building



Opportunities, Challenges and Risks

● آقای هوش مصنوعی ●

رسانه هوش مصنوعی دانشگاه تهران

Benjamin Schaller

@MrArtificialintelligence

Abstract This chapter explores the opportunities, challenges and risks of using artificial intelligence (AI) technologies in the context of conventional arms control and military confidence-building. First, it briefly reflects upon different theoretical approaches and perspectives on arms control and military confidence-building. Second, it provides a brief overview of existing treaties, regimes and measures in Europe. Finally, the chapter concludes with a few reflections and food for thought on the opportunities, challenges and risks inherent in AI technologies for: (1) the balance of power; (2) analysis, planning, coordination, and evaluation; (3) verification and (4) trust-building in conventional arms control and military confidence-building measures in Europe. In sum, the chapter argues that the best prospects for AI technologies in conventional arms control and military confidence-building are in the augmentation of human intelligence, while the biggest risks lie in a lack of human oversight and an uncritical reliance on AI systems, as well as in the reduction of the trust-building effects of direct military-to-military contacts.

1 Introduction

The recent deterioration of relations between the North Atlantic Treaty Organization (NATO) and Russia has put conventional arms control and military confidence-building back on the agenda of defense and security-policy discussions in Europe. For many years, this hardening of already opposing positions on the role and future direction of conventional arms control and military confidence-building in Europe has blocked serious modernization efforts. Still largely reflecting the political, military and technological realities of the end of the Cold War, conventional arms control and confidence- and security-building measures (CSBM) have for many years been struggling with a constantly diminishing role in European defense and security politics. While the global order has witnessed an increasing shift toward

B. Schaller (✉)

UiT - The Arctic University of Norway, Tromsø, Norway

e-mail: bsc009@post.uit.no

multipolarity (e.g., through the rise of new major powers, such as China), participating states of the Organization for Security and Co-operation (OSCE) are embroiled in debate over alleged cases of non-compliance and political disagreements on the role and future direction of conventional arms control in Europe, turning the OSCE into another arena where the political and strategic tensions and disagreements between Russia and the West unfold (Charap et al., 2020, pp. 1–2; Koivula, 2017, p. 113; Schaller, 2020, pp. 126–128).

In such a political gridlock situation, there has been little room for discussion of the potential role and impact that artificial intelligence (AI) technologies could have in the modernization and future of conventional arms control and military confidence-building in Europe.¹ Consequently, in the absence of a broader academic or political debate, this chapter will approach the topic by discussing the opportunities, challenges and risks of AI technologies in conventional arms control and military confidence-building in Europe from a more practical/conceptual as well as theoretical point of view. To this end, the chapter is divided into three sections:

First, I will reflect upon previous theoretical approaches to the role of arms control and military confidence-building in defense and security politics. Secondly, I will provide a brief overview of central treaties and documents as well as current challenges faced by conventional arms control and military confidence-building in Europe. Finally, I conclude with some initial reflections about the potential opportunities, challenges and risks of AI technologies in conventional arms control and military confidence-building. More specifically, I will reflect upon their potential role and impact on

- existing categories of military forces and equipment,
- the planning and coordination,
- implementation, and
- analysis and evaluation of existing regimes.

2 Theoretical Approaches to Arms Control and Military Confidence-Building

This section provides a brief overview of existing theoretical approaches to international relations that—consciously or unconsciously—shape the perspectives of scholars, policymakers and practitioners on the role of arms control and military confidence-building in defense and security policy. These perspectives will also very probably influence their positions regarding the introduction of AI technologies into conventional arms control and military confidence-building in Europe.

¹This does not refer to the numerous discussions and debates presented in this book, but to debates specifically focusing on conventional arms control and military confidence-building in the context of the OSCE.

Looking at previous academic debates, it seems fair to conclude that, apart from a small expert community, arms control and military confidence-building have suffered from a considerable decline in interest in defense and security-policy debates since the end of the Cold War. This waning interest probably also explains why some of the most influential theoretical approaches date back to a period between the 1960s and the 1990s (e.g. Bull, 1961; Darilek, 1992; Schelling & Halperin, 1961). More recent academic debates are dominated by empirical case studies and policy analysis and primarily focus on the role of arms control and military confidence-building in relations among the United States (US), NATO and the Russian Federation (or previously the Soviet Union) (e.g. Koivula & Simonen, 2017; Kühn, 2013; Lachowski, 2004). The theoretical underpinnings of most studies can be grouped loosely into three different theoretical camps in international relations theory: structural realism, neoliberal institutionalism and constructivism.

2.1 The Structural Realist Approach to Arms Control and Military Confidence-Building

The first group has its roots in a more traditional *structural realist* understanding of international relations that is primarily concerned with the constant struggle for survival faced by states in the anarchical structure of the international system (e.g. Jervis, 1978; Mearsheimer, 2014; Waltz, 1979). Consequently, scholars in this tradition look at arms control and CSBM primarily through the lens of “balances of power,” emphasizing measures that reduce, limit or impose constraints on the military capabilities of states and arguing for highly stringent and comprehensive verification regimes (e.g. Bull, 1961; Peters, 2000; Schelling & Halperin, 1961).

2.2 The Neoliberal Institutional Approach to Arms Control and Military Confidence-Building

The second camp is rooted in *neoliberal institutionalism*, which holds that the anarchy in the international system can be overcome by the establishment of international institutions, norms and laws that guide the behavior and facilitate cooperation among states (e.g. Keohane, 1984; Keohane & Nye, 1999). Scholars in this tradition assess the ability of arms control and CSBM to address and overcome the problems of anarchy—most notably the security dilemma—in defense and security relations among states by emphasizing the importance of increased transparency and predictability in connection with military forces, capabilities and activities, such as through regular exchanges of information and credible verification mechanisms that reduce the risk of deception (e.g. Borawski, 1986; Darilek, 1992; Vick, 1988).



Fig. 1 Theoretical approaches to arms control and military confidence-building. Own illustration

2.3 *The Constructivist Approach to Arms Control and Military Confidence-Building*

The third and probably least influential camp in scholarly debates on arms control and CSBM has its roots in a *constructivist* understanding of international relations. Constructivist scholars rebut traditional arguments that anarchy and security dilemmas are inherent features of the international system, arguing instead that both are constructed through the behavior and interactions among states (e.g. Finnemore, 1996; Wendt, 1992). Constructivist scholars usually focus on the ability of arms control and CSBM to facilitate direct contacts and cooperative approaches to security, which are seen as facilitating the formation of more trusting defense and security relations among states (e.g. Adler, 1998, p. 128; Schaller, 2020, p. 24).

2.4 *Summary*

The three different theoretical camps are summarized in the subsequent overview (Fig. 1).

3 **Conventional Arms Control and Military Confidence-Building in Europe: A Brief Overview**

The current framework of interlocking arms control and CSBM regimes in Europe dates back to a period of rapprochement that led to the adoption of a number of central documents and treaties that culminated in the post-Cold War European

security architecture (e.g. the Helsinki Final Act of 1975, the NATO-Russia Founding Act of 1997, the Paris Charter of 1990). Each treaty and document was meant to help bridge the divide that had separated East and West for many years. This political process was also accompanied by a number of military negotiations, such as the Mutual and Balanced Force Reductions (MBFR) talks between NATO and the Warsaw Pact (Goldblat, 2002, pp. 220–222), negotiations about military confidence-building measures in the context of the Conference on Security and Co-operation in Europe (CSCE) (Goldblat, 2002, pp. 257–260), as well as consultations on the establishment of an aerial observation regime. These negotiations and talks led to the adoption of the Vienna Document on Confidence- and Security-Building Measures (VDoc), the Treaty on Conventional Armed Forces in Europe (CFE) in 1990 and the Treaty on Open Skies (OS) in 1992.

3.1 The Vienna Document (VDoc)

The VDoc was specifically designed to reduce the risk of surprise attacks and unintended escalation between OSCE participating States by increasing transparency in connection with military forces and their activities in Europe. This was largely achieved by means of a comprehensive annual exchange of military information (e.g. on military forces, major weapon and equipment systems, or defense planning) (VDoc, 2011, Ch. I & II), the prior announcement and observation of larger military exercises and activities (VDoc, 2011, Ch. V & VI), as well as the possibility of conducting a small number of inspections and evaluation visits in order to verify samples of the military information that is regularly provided under the document (VDoc, 2011, Ch. IX). With its specific focus on formation of trust, the document also facilitates various opportunities for regular contacts between militaries from OSCE participating States (e.g. visits to air bases and other military facilities, demonstrations of new weapon systems, seminars) (VDoc, 2011, Ch. IV). Since its first adoption, the VDoc has been updated and modernized four times, most recently through minor technical adaptations in 2011. A scheduled update of the document in 2016 was prevented by Russia, which stated that NATO's "policy of military containment of Russia and the Alliance's concrete steps in the military sphere rule out the possibility of reaching agreements on confidence-building" (Forum for Security Co-operation [FSC], 2016, Annex 3).

3.2 The Treaty on Conventional Armed Forces in Europe (CFE Treaty)

The CFE Treaty is a full-fledged conventional disarmament and arms control treaty, originally designed to reduce the possibility of major offensive military operations

and to contribute to a stable military balance at a lower level of conventional military forces between NATO and the (former) Warsaw Pact. To this end, in contrast to the VDoc, the CFE Treaty defines total and regional limits for holding and deployment of five major conventional weapon systems in Europe, namely battle tanks, armored combat vehicles, artillery, combat aircraft and combat helicopters (CFE Treaty, Art. IV, V & VI). In addition, the CFE Treaty established a very detailed and comprehensive system of annual exchanges of information and regular notification of changes in the conventional armed forces of states parties (e.g. about the command structure, total holdings, personnel, dislocation sites, entry and exit into the area of application, and uncommissioned equipment) (CFE treaty, Art. XIII & Protocol on Notification and Exchange of Information). The provisions of the CFE Treaty are complemented by a particularly thorough and comprehensive verification regime that allows states parties to more credibly monitor and verify the compliance of all parties with the treaty's provisions (e.g. through a considerably higher number of inspection quotas or more rights for inspection teams) (CFE Treaty, Art. XIV & Protocol on Inspection). In the course of the treaty, which only applies to NATO member states prior to the end of the Cold War as well as the successor states of the former Warsaw Pact, approximately 60,000 heavy weapon systems in Europe have been successfully destroyed (Federal Foreign Office, 2018). Due to a dispute between Russia and NATO states over the full withdrawal of Russian troops from Georgia and Moldova ("Istanbul Agreements"), the ratification of the Adapted Treaty on Conventional Armed Forces in Europe (ACFE)—a modernized version of the CFE Treaty—failed. This failure led Russia to unilaterally stop its implementation of the CFE Treaty in 2007 and to withdraw from the treaty's Joint Consultative Commission in 2015, resulting in a political deadlock that remains unresolved to this day (see Koivula, 2017, p. 120; Mankoff, 2012, pp. 130–131; North Atlantic Treaty Organization [NATO], 2018).

3.3 *The Treaty on Open Skies (OS)*

The treaty on OS is a military transparency and confidence-building regime that allows its 34 signatory states to conduct observation flights over the territories of all treaty states. To this end, the treaty defines rules for quota distribution (Treaty on OS, Art. III), defines flight procedures and mission planning (Treaty on OS, Art. VI) and specifies technical details for observation aircraft, cameras and sensors (Treaty on OS, Art. IV & V). The treaty does not have its own dedicated system of military information exchange, but primarily relies on the information provided through other treaties and documents, such as the VDoc and the CFE treaty. In addition, the Treaty on OS is the only conventional arms control and CSBM regime in Europe that also covers the land territories of Canada, the United States and Russia (east of the Ural Mountains). It contains strict regulations regarding the cameras and sensors used (e.g. limits on the maximum ground resolution) (Treaty on OS, Art. IV), is cooperative in nature (flights are conducted in cooperation with the inspected state)

(Treaty on OS, Art. VI) and all treaty states have the right to acquire images taken during OS missions (Treaty on OS, Art. IX). While the treaty has also been used during the early stages of the crisis in and around Ukraine (Bell & Wier, 2019), its main focus is more on military confidence-building, than military surveillance. The treaty on OS has also been negatively affected by deteriorating relations between Russia and the West. This has given rise to a number of implementation disputes, most notably between Russia and the United States, such as restrictions on the maximum flight distance over Kaliningrad or regarding observation flights in proximity to the Russian-Georgian border (Bell et al., 2020, pp. 2–3; Bell & Wier, 2019). In 2020, the dispute between Russia and the United States culminated in the United States withdrawing from the treaty completely, (United States Department of State, 2020) with Russia following the end of 2021 (TASS, 2021). In short, the treaty on OS has—just like other arms control and CSBM regimes—come under serious political pressure in recent years.

3.4 Major Challenges

Western-Russian relations have been in a slow but constant decay for many years and their drastic deterioration since 2014 has only reinforced the already entrenched positions on both sides. Having missed various opportunities for adapting existing measures and regimes to the current political, military and technological realities, conventional arms control and CSBM are struggling to fulfill their stabilizing and trust-building function in the European security architecture. Some of the most notable shortcomings in this regard are:

- Gaps and loopholes in existing provisions and regulations (e.g. excessively high thresholds for the notification or observation of exercises and activities, or possibilities for conducting large-scale exercises without prior notification or observation);
- Insufficient inclusion of naval and paramilitary forces as well as of new technologies and weapon systems (e.g. AI, drones or autonomous weapon systems);
- Developments in military structures, doctrines and strategies (e.g. shifts toward smaller, highly deployable forces with considerable firepower); as well as
- Changes in conflict scenarios (e.g. the increasing number of intrastate conflicts and hybrid/gray-zone attacks) (Charap et al., 2020; Koivula, 2017; Schaller, 2018).

4 Artificial Intelligence and Machine Learning in Conventional Arms Control: Opportunities, Challenges and Risks

This section will present initial reflections about the potential opportunities, challenges and risks of AI technologies in conventional arms control and military confidence-building. More specifically, it will reflect upon how AI technologies may affect the main principles and mechanisms of arms control and CSBM and how they may be able to support the implementation of these mechanisms. Specific attention will be paid to the challenges and risks of *machine learning*, *data mining*, *visual analytics* and *augmented intelligence*. On the basis of the previous two sections of this chapter, four main questions regarding the potential role, impact and opportunities inherent in these challenges and risks can be identified:

1. To what extent will AI technologies affect the balance of power between states?
2. How can AI technologies help in the analysis of information about military capabilities and activities?
3. To what extent can AI technologies contribute to more thorough and comprehensive verification?
4. How will the increasing use of AI technologies affect the interaction and frequency of direct military-to-military contacts on the ground?

Each question is discussed in greater detail in the four following sections.

4.1 *Changes in the Balance of Power*

Advances in AI technologies, such as machine learning (ML), will challenge the already delicate and somewhat outdated “balance of power” that many traditional arms control regimes seek to establish or maintain between political and military opponents. One of the issues that is currently the subject of probably the most controversy is the potential control of so-called “Lethal Autonomous Weapon Systems” (LAWS) under the United Nations Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons (CCW) (e.g. Friman, 2017; Koivula, 2017, p. 125; Maas, 2019; also see the text by Anka Dahlmann in this volume).

Apart from increasing autonomy, emerging AI technologies will also have a considerable impact on the processing power, precision and military application of existing weapon and equipment systems. This considerably reduces the reaction time for decision-makers and may undermine the existing balances of power among states. In this regard, advances in AI technologies seem to reinforce the arguments of those who call for an increased focus on qualitative factors in conventional arms control and CSBM. While the focus has traditionally been on numerical limits for certain weapon categories and systems (e.g. battle tanks) establishing stable balances

of power, today there is a need for a more qualitative approach to arms control that also looks at the capacities of different weapon and equipment systems (e.g. mobility, (re)deployment, fire power) (Koivula, 2017, pp. 123–127; Kühn, 2013, pp. 196–198; Schaller, 2018, pp. 116–117). In this context, this raises the question of how viable and comparable purely numerical limits on weapon systems and equipment are, if their capacities can be significantly amplified by advanced sensors and software solutions. For example, AI technologies, such as ML or visual analytics, could support the identification of enemy targets and speed up human decision-making processes or make decisions to engage with certain targets completely on their own. Considering that such qualitative differences are hard to compare and that software codes may have different error rates and can be altered rather quickly, determining and verifying whether a certain weapon system is capable of performing such tasks, is not only of grave ethical concern, but also poses significant challenges to traditional verification measures, during which specially trained verification officers conduct on-site inspections of military equipment and personnel.

Nevertheless, within the OSCE, the role and impact of AI technologies has, so far, only played a minor role. The current problems and positions of different stakeholders in arms control and CSBM (e.g. allegations of non-compliance, the role and impact of regional conflicts in the OSCE area) are too great and too strongly entrenched. However, sooner or later, the technological advances in the field of AI will make discussions on the future of conventional arms control and CSBM in Europe. This will be especially true for the necessity of including new weapon systems, such as drones and automated systems, in existing as well as future documents and regimes.

4.2 Analysis, Planning, Coordination and Evaluation

AI technologies will most probably also change the way countries analyze the military information that is exchanged through different arms control treaties and regimes as well as the way they plan, coordinate, and evaluate their arms control and CSBM activities (e.g. verification measures).

Already today, countries rely on AI technologies to analyze and detect anomalies and trends in particularly large and complex data sets (“big data”). The information gathered through such systems is then often used for risk assessment, crisis detection and early warning. For example, the German Federal Foreign Office uses ML for detecting potential crisis situations at as early a stage as possible (see Federal Foreign Office, 2020), while data scientists have developed advanced systems for disaster management and crisis response after natural disasters (e.g. Mittelstädt et al., 2015; Wang et al., 2018). While these systems are not free of errors or potential biases—they detect only what they are programmed to detect—with the further development AI technologies (e.g. in error reduction, image or pattern recognition), the role and importance of such systems will further increase in the future.

In addition, in the context of arms control and CSBM, states exchange large amounts of military information. For example, under the Vienna Document's Annual Exchange of Military Information (AEMI), participating states exchange information about the organization, size and location of their troops as well as the size and location of major weapon and equipment systems (e.g. battle tanks, artillery systems, armored vehicles) (VDoc, 2011, Ch. I & II). Similarly, even more detailed exchanges of military information also take place under the provisions of the CFE Treaty (CFE Treaty, Art. XIII & Protocol on Notification and Exchange of Information). These exchanges consist of long lists and tabular data, listing equipment, troops and their locations.² Digitalized, exploratory data analysis with data mining, visual analytics, and ML can assist arms control units in the analysis and monitoring of this information. For example, at their best AI solutions monitored by trained and experienced personnel can help arms control officers and governments and improve the detection of trends, developments and changes in the force structure, activities, capabilities and deployment of OSCE participating states (e.g. through visual processing or data mining). The findings of such processes could then be fed into different (national and multinational) early-warning mechanisms and assist arms control units and governments in target selection for inspections and on-site verification activities. In addition, AI systems could also keep track of other countries' verification activities, for example, when a certain site was inspected for the last time, what the findings of that mission were, and whether any significant changes have been reported since.³ AI systems can also assist arms control units in post-inspection evaluation, such as in the comparison of findings with information that is exchanged and the results of previous verification measures, or in the processing and analysis of images under the Treaty on Open Skies. For example, using pattern recognition AI technologies could learn the typical strategic elements and infrastructure of military bases and airfields and support the detection of changes in their setup.

However, considering the obvious limitations of AI technologies (e.g. their capacity and error rate depend on their programming and training), it is important that AI solutions continue to be combined with human expertise, experience and instinct. The main benefit of such "human-in-the-loop hybrid-augmented intelligence" systems—systems that rely on a combination of human intelligence, enhanced by technical means, such as visual analytics, ML or data mining (see Zheng et al., 2017, p. 154)—lies in their ability to assist experienced and well-trained arms control officers in the development of recommended actions for practitioners and policymakers. This can ensure faster, more informed, and ideally more appropriate government reactions (e.g. in emerging crisis situations), while practitioners receive more time for direct interaction with arms control officers from other countries, a key aspect of military confidence-building (Schaller, 2020, pp. 236–238).

²This can, for instance, be seen in the publicly available Annual Exchange of Military Information (AEMI) of the Finnish armed forces (see The Finnish Defence Forces, 2019).

³Some countries are already using comparable systems and software solutions.

4.3 *More Thorough and Comprehensive Verification*

AI technologies can also make a significant contribution to the verification of arms control and CSBM regimes. More specifically, at their best they can facilitate more thorough and comprehensive verification that can be implemented whenever the deployment of inspection personnel would be too dangerous (e.g. in emerging crisis situations) or politically difficult (e.g. due to status implications in contested regional conflicts).

Technical means of verification (e.g. infrared, radar or telemetric sensors) tend to retain the upper hand under difficult environmental and light conditions (e.g. rain, twilight, clouds, fog) (Goldblat, 2002, pp. 314–315). Cameras, sensors, and the software underlying them could even be trained not only to (self-)identify weapon systems and military equipment during verification measures (on the ground or in the air), but also to detect possible variations and modifications (e.g. new antennas or setups). In combination with well-trained and experienced personnel, the quality of such technical means of verification could be even further enhanced. For example, through credible monitoring and training, such augmented intelligence systems could considerably reduce the risk of deception (e.g. through additional structures on vehicles or lookalikes), allowing for a particularly robust verification regime.

In addition, AI technologies could assist or, where necessary, completely replace national verification teams in situations where deployment of military personnel and experts on the ground would be too dangerous (such as conflict zones) or politically sensitive (such as potential status implications in protracted conflict situations). For example, during emerging crisis situations, technical means of verification (e.g. drones, CCTV, sensors) could be made available to a neutral third party (e.g. the OSCE Conflict Prevention Centre), as it was the case for the OSCE Special Monitoring Mission to Ukraine (OSCE Special Monitoring Mission to Ukraine, 2019). Complemented by AI technologies, these technical means of verification could be used to provide a comprehensive, fast and impartial picture of the security situation, allowing OSCE participating states to react more swiftly and appropriately to a quickly evolving crisis situation. For example, as mentioned above, AI technologies could be used to (self-)identify weapon systems and military equipment on the ground or to ensure a swifter analysis and evaluation of verification findings. In addition, augmented intelligence systems could also be used to monitor agreements between participating states and non-recognized entities in protracted conflict situations (e.g. Abkhazia, South Ossetia or Transnistria). Whereas the deployment of national verification personnel would have considerable status implications for these entities under international law (Kapanadze et al., 2017, p. 12), technical means of verification could offer possible workarounds. For example, the monitoring of trade between Russia and Georgia is carried out by a private company, which *inter alia* makes use of electronic seals and GPS tracking, a solution that was necessary as the two sides could not agree on how to regulate trade involving the Georgian break-away territories of South Ossetia and Abkhazia (Civil.ge, 2011).

While technologically feasible, the main challenges for such more technical verification regimes are to establish sufficient levels of trust in the technical solutions provided as well as the predictably difficult negotiations between states on granting each other or a neutral third party such far-reaching and intrusive verification possibilities. Such a precedence exists, in particular in the area of nuclear arms control, such as satellite-based monitoring of missile launchers (Goldblat, 2002, p. 315) or in the form of technical (such as CCTV) and on-site monitoring of nuclear safeguards by the International Atomic Energy Agency (IAEA) (Goldblat, 2002, pp. 318–319). However, the use of technical equipment (e.g. digital cameras or satellite positioning systems) during conventional arms control inspections is often met with skepticism and notable national reservations. Within the context of the OSCE, this can be seen in previous disagreements over the certification of digital cameras, infrared sensors or radar sensors under the Treaty on Open Skies (Bell & Wier, 2019). In fact, even today, most aircraft still fly with old analog cameras and sensors, while the transition to digital cameras is just getting under way.

4.4 The Importance of Maintaining the “Human Factor” in Arms Control

At their best, AI technologies potentially offer many opportunities for increasing the level of transparency and predictability in defense and security relations between states. However, they also pose significant risks and challenges that underline the importance of preserving the essential “human factor” in military confidence-building and arms control.

First, due to the fact that the performance and reliability of AI systems depends to a considerable extent on their programming and training, which—as already discussed above—can lead to errors and unintended biases, there is merit in focusing on the development of “human-in-the-loop” solutions. Such “human-in-the-loop” systems can also help overcome the “difficult question of liability and trust of governments to base their own policies on AI systems, with an—at least for them—often opaque decision-making processes” (“black box” character). Here, arms control officers and specifically trained personnel will play an important role in interpreting, explaining and translating the findings of AI systems into viable recommendations for political decision-makers.

In addition to the difficult questions of bias, liability and mistrust, there is also the risk of AI technologies significantly reducing the amount of direct contact and cooperation among human personnel. While this *human factor* in arms control and CSBM usually receives less attention in academia and politics, numerous historical examples, personal experiences and reports of arms control officers suggest that individual judgment, professional intuition and the interpersonal level of trust in arms control and CSBM play a crucial role and should be actively safeguarded as an essential component in arms control and military confidence-building (Lewis et al.,

2014, pp. 24–27; Schaller, 2020, pp. 168–180; Welch Larson, 1997, pp. 713–717). As Allan S. Krass put it regarding the interplay of verification and trust in arms control: “some initial trust must be present if any verification system is to succeed in preserving an arms control or disarmament agreement” (Krass, 1985, p. 287).

This level of professional and interpersonal trust appears to be particularly valuable in times of increased political or military tension when other channels of exchange and interaction are either strained or shut down. For example, since the beginning of the crisis in and around Ukraine, the VDoc and the Open Skies Treaty offer some of the few remaining venues for facilitating direct contacts between Russian and Western military officers (Schaller, 2020, pp. 58–60). Both the post-Cold War period and constructivist approaches to IR emphasize that such personal contacts and interactions are absolutely indispensable for the development of more trusting defense and security relations between states (e.g. Adler, 1998, p. 128; Schaller, 2020, p. 24). Consequently, with increased advances in AI technologies it is important that governments find ways of maintaining a credible level of direct interaction and exchange among arms control officers on the ground. Increasing use of AI technologies in the implementation of conventional arms control and CSBM regimes in Europe should not result in even further cuts in budgets and personnel that have already led to a considerable loss of expertise and have undermined their trust-building effects. Ideally, these technologies should augment the capacities of arms control officers, relieving them of routine and time-consuming tasks so they can spend more time on their interactions and direct exchanges with colleagues from different OSCE participating states.

In short, AI technologies should not be replacing, but instead assisting well-trained and experienced personnel by means of augmented intelligence solutions that allow a sufficient level of human oversight and military-to-military interaction in conventional arms control and CSBM.

5 Concluding Remarks

While contrasting views and renewed tensions between Russia and the West have led to a standstill in the discussions on the role and future direction of conventional arms control and military confidence-building in Europe, technological advances, also in the field of artificial intelligence, will leave their mark on the European security landscape and affect the functioning of existing as well as future arms control and CSBM regimes. In the absence of a more serious political and academic discourse, this chapter discussed the opportunities, challenges and risks of AI and ML for conventional arms control and military confidence-building in Europe. More specifically, the chapter has highlighted how AI technologies may affect the military balance of power between states, but also how they may in the best case also support the work of arms control officers. For example, if operating reliably and in an unbiased way, AI technologies could be used to detect anomalies in the regular exchanges of military information, to help arms control units plan and coordinate

verification activities, and to analyze and evaluate data and verification findings faster and more accurately. Together with the potential use of technical means of verification (e.g. drones, CCTV) in difficult political situations or where the deployment of verification personnel would be too dangerous, AI technologies have considerable potential for offering a swift and more impartial picture of the security situation on the ground, so that governments can react quickly and with greater precision to emerging crisis situations. However, despite these potential benefits, it is important that AI technologies, which are prone to errors and biases, remain in a supporting and supplementary role in conventional arms control and military confidence-building. Direct military-to-military contacts and cooperative approaches to security are too important in the formation of trust in defense and security relations between states to be replaced by AI, just as careful human assessment as well as the development of well-considered responses and recommended actions are too sensitive and important for political decision-makers for them to be abandoned. In short, the greatest opportunity for AI technologies in conventional arms control and CSBM in Europe lies in the augmentation, not in the replacement of human intelligence.

Acknowledgments I would particularly like to thank a very good friend of mine for his extremely useful input, feedback and advice regarding the technical aspects, elements and status quo of AI technologies. Thank you very much!

References

- Adler, E. (1998). Seeds of peaceful change: The OSCE's security community-building model. In E. Adler & M. N. Barnett (Eds.), *Security communities* (pp. 119–160). Cambridge University Press.
- Bell, A., Richter, W., & Zagorski, A. (2020). *How to fix, preserve and strengthen the Open Skies Treaty* [Issue Brief No. 9]. The Deep Cuts Commission. Retrieved from https://deepcuts.org/files/pdf/Deep_Cuts_Issue_Brief_9-Open_Skies_Treaty.pdf
- Bell, A., & Wier, A. (2019). *Open skies treaty: A quiet legacy under threat*. Arms Control Association. Retrieved from <https://www.armscontrol.org/act/2019-01/features/open-skies-treaty-quiet-legacy-under-threat>
- Borawski, J. (1986). Confidence-building measures. Rescuing arms control. *The Fletcher Forum of World Affairs*, 10(1), 111–131.
- Bull, H. (1961). *The control of the arms race: Disarmament and arms control in the missile age*. Praeger.
- Charap, S., Lynch, A., Drennan, J. J., Massicot, D., & Persi Paoli, G. (2020). *A new approach to conventional arms control in Europe: Addressing the security challenges of the 21st century*. Research Report. RAND Corporation. Retrieved from https://www.rand.org/pubs/research_reports/RR4346.html
- Civil.ge. (2011). Georgia-Russia WTO deal in details. *civil.ge*. Retrieved from <https://civil.ge/archives/121546>
- Darilek, R. E. (1992). The theory of confidence-building measures. In J. E. Naton (Ed.), *The de-escalation of nuclear crises* (pp. 3–35). Palgrave Macmillan.

- Federal Foreign Office. (2018). *Treaty on conventional armed forces in Europe (CFE Treaty)*. Retrieved November 22, 2018, from <https://www.auswaertiges-amt.de/en/aussenpolitik/themen/abruestung/uebersicht-konvalles-node/ruestungskontrolle/kse-vertrag-node>
- Federal Foreign Office. (2020). *Krisenfrüherkennung, Konfliktanalyse und Strategische Vorausschau*. Retrieved from <https://www.auswaertiges-amt.de/de/aussenpolitik/themen/krisenpraevention/-/2238138>
- Finnemore, M. (1996). *National interests in international society*. Cornell University Press.
- Finnish Defence Forces. (2019, November 26). *Annual exchange of military information 2020—Finland*. Retrieved from https://puolustusvoimat.fi/documents/1948673/2015391/FI_AEMI_2020.pdf/a5b066cf-cd1f-7fec-5213-35fc67b70ae8/FI_AEMI_2020.pdf
- Forum for Security Co-operation. (2016, November 9). *Special meeting of the forum for security co-operation (834th plenary meeting)*. OSCE.
- Friman, J. (2017). The Pandora's box of military artificial intelligence. In T. Koivula, & K. Simonen (Eds.), *Arms control in Europe. Regimes, trends and threats* (National Defence University Series 1, Research publications, No. 16, pp. 133–150). National Defence University.
- Goldblat, J. (2002). *Arms control. The new guide to negotiations and agreements*. Sage.
- Jervis, R. (1978). Cooperation under the security dilemma. *World Politics*, 30(2), 167–214.
- Kapanadze, S., Kühn, U., Richter, W., & Zellner, W. (2017, January). *Status-neutral security, confidence-building and arms control measures in the Georgian context*. CORE Working Papers no. 28. Institute for Peace Research and Security Policy.
- Keohane, R. O. (1984). *After hegemony: Cooperation and discord in the world political economy*. Princeton University Press.
- Keohane, R. O., & Nye, J. S. (1999). *Power and Interdependence* (2nd ed.). Longman.
- Koivula, T. (2017). Conventional arms control in Europe and its current challenges. In T. Koivula & K. Simonen (Eds.), *Arms control in Europe. Regimes, trends and threats* (National Defence University Series 1, Research publications, No. 16, pp. 113–132). National Defence University.
- Koivula, T., & Simonen, K. (Eds.). (2017). *Arms control in Europe. Regimes, trends and threats* (National Defence University Series 1, Research publications, No. 16). National Defence University.
- Krass, A. S. (1985). Verification and trust in arms control. *Journal of Peace Research*, 22(4), 285–288.
- Kühn, U. (2013). Conventional arms control 2.0. *The Journal of Slavic Military Studies*, 26(2), 189–202. <https://doi.org/10.1080/13518046.2013.779859>
- Lachowski, Z. (2004). *Confidence- and security-building measures in the New Europe*. SIPRI Research Report no. 18. Stockholm International Peace Research Institute.
- Lewis, P., Williams, H., Pelopidas, B., & Aghlani, S. (2014, April). *Too close for comfort. Cases of near nuclear use and options for policy*. Chatham House Report. The Royal Institute of International Affairs.
- Maas, M. M. (2019). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy*, 40(3), 285–311. <https://doi.org/10.1080/13523260.2019.1576464>
- Mankoff, J. (2012). *Russian foreign policy. The return of great power politics* (2nd ed.). Rowman & Littlefield.
- Mearsheimer, J. J. (2014). *The tragedy of great power politics*. Norton.
- Mittelstädt, S., Wang, X., Eaglin, T., Thom, D., Keim, D., Tolone, W., & Ribarsky, W. (2015). An integrated in-situ approach to impacts from natural disasters on critical infrastructures. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 48, 1118–1127.
- North Atlantic Treaty Organization. (2018). *NATO's role in conventional arms control*. North Atlantic Treaty Organization. Retrieved July 6, 2019, from https://www.nato.int/cps/ru/natohq/topics_48896.htm?selectedLocale=en
- OSCE Special Monitoring Mission to Ukraine. (2019). *OSCE SMM technical monitoring*. Retrieved from <https://www.osce.org/special-monitoring-mission-to-ukraine/419582>

- Peters, J. E. (2000). *The changing quality of stability in Europe. The conventional forces in Europe treaty toward 2001*. RAND Corporation.
- Schaller, B. (2018). Back to the future? Revisiting military confidence-building in Europe*. *Sicherheit und Frieden*, 36(3), 115–120. <https://doi.org/10.5771/0175-274X-2018-3-115>
- Schaller, B. (2020). *Trust and distrust in defence & security politics. A multi-level analysis of the defence and security relations between Norway, Sweden, Canada, and Russia*. Doctoral dissertation. The Arctic University of Norway, Tromsø.
- Schelling, T. C., & Halperin, M. H. (1961). *Strategy and arms control*. Twentieth Century Fund.
- TASS. (2021, May 5). *Russian government approves proposal to denounce Open Skies Treaty*. Retrieved from <https://tass.com/politics/1286701>
- Treaty on Conventional Armed Forces*, Paris, November 19, 1990.
- Treaty on Open Skies*, Helsinki, 1 January, 2002. Retrieved from <https://www.osce.org/files/f/documents/1/5/14127.pdf>
- United States Department of State. (2020, November 22). *Treaty on open skies* [press release]. Retrieved from <https://2017-2021.state.gov/treaty-on-open-skies/index.html>
- VDoc. (2011). *Vienna document 2011 on confidence- and security-building measures*. Retrieved from Organization for Security and Co-operation in Europe website: <https://www.osce.org/files/f/documents/a/4/86597.pdf>
- Vick, A. J. (1988, March). *Building confidence during peace and war* (RAND Corporation, Ed.) (A RAND Note). RAND Corporation.
- Waltz, K. N. (1979). *Theory of international politics*. McGraw-Hill.
- Wang, R.-Q., Mao, H., Wang, Y., Rae, C., & Shaw, W. (2018). Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Computers and Geosciences*, 111, 139–147. <https://doi.org/10.1016/j.cageo.2017.11.008>
- Welch Larson, D. (1997). Trust and missed opportunities in international relations. *Political Psychology*, 18(3), 701–734.
- Wendt, A. (1992). Anarchy is what states make of it: The social construction of power politics. *International Organization*, 46(2), 391–425.
- Zheng, N., Liu, Z., Ren, P., Ma, Y., Chen, S., Yu, S., et al. (2017). Hybrid-augmented intelligence: Collaboration and cognition. *Frontiers of Information Technology and Electronic Engineering*, 18(2), 153–179. <https://doi.org/10.1631/FITEE.1700053>

Cyber Weapons and Artificial Intelligence: Impact, Influence and the Challenges for Arms Control



● آقای هوش مصنوعی ●

Thomas Reinhold and Christian Reuter

رسانه هوش مصنوعی دانشگاه تهران

@MrArtificialintelligence

Abstract As cyber weapons and artificial intelligence technologies share the same technological foundation of bits and bytes, there is a strong trend of connecting both, thus addressing the imminent challenge of cyber weapons of processing, filtering and aggregating huge amounts of digital data in real time into decisions and actions. This chapter will analyze this development and highlight the increasing tendency towards AI enabled autonomous decisions in defensive as well as offensive cyber weapons, the arising additional challenges for attributing cyberattacks and the problems for developing arms control measures for this “technology fusion”. However, the article also ventures an outlook how AI methods can help to mitigate these challenges if applied for arms control measures itself.

1 Introduction

The idea of the weaponization of cyber tools has been under discussion for some time (Reinhold & Reuter, 2019b; Werkner & Schörnig, 2019). Many military or national security doctrines worldwide have adapted to the development that software can be designed, injected, triggered and controlled in foreign IT systems to perform tasks ranging from espionage to sabotage. This has been done from the perspective of necessary and appropriate defensive measures but also partly as a new category for offensive planning. Although no common international understanding has yet been reached on the threats posed by cyber weapons and their prevention, let alone a binding legal instrument, this field is already beginning to change due to the emergence of improved algorithms in artificial intelligence and machine learning (AI/ML) and their potential application for or against cyber weapons (Schörnig, 2018; US-DOD, 2018b). Given the fact that cyber and AI/ML measures are *natural siblings* from a technical perspective, the following text provides an assessment of

T. Reinhold (✉) · C. Reuter

Chair of Science and Technology for Peace and Security (PEASEC), Department of Computer Science, Technical University of Darmstadt, Darmstadt, Germany

e-mail: reinhold@peasec.tu-darmstadt.de; reuter@peasec.tu-darmstadt.de

how AI/ML methods could influence the development of malicious cyber activities based on an overview of their current state. Regarding the threats posed by this development for international security and new challenges for arms control, the text seeks on the one hand to assess how arms control approaches should prepare for AI/ML-driven cyber weapons. On the other hand, the text also examines the question of whether and how this technology can improve arms control approaches combating the weaponization of cyberspace.

2 Cyber Weapons and the Militarization of Cyberspace

Technological and scientific advances, especially the rapid evolution of information technology (IT), play a crucial role in questions of peace and security (Reuter, 2019). First and foremost, the most significant impact of the discussions and developments regarding the weaponization of cyberspace in recent years has been on its influence and the changes it has introduced to national and international security doctrines. An important incident has been the discovery of Stuxnet (Langner, 2013), malware developed by the US and Israel (Nakashima & Warrick, 2012) and targeted against a specific nuclear enrichment facility in Iran. Stuxnet manipulated the industrial control system of the facility by covertly changing thresholds and parameters of the control software to sabotage the enrichment process. This highly specified and *hand-crafted* attack on IT systems forced state leaders and decision-makers to recognize the vulnerabilities in computer systems and the threat that arises from the high degree of dependency on IT in economic, societal and government sectors. Especially critical infrastructures are now perceived to be high-risk targets for state and non-state cyberattacks. Although this was not the first cyber incident, and was hardly news for IT security specialists, the Stuxnet event demonstrated the technological possibility of crossing the cyber-physical barrier with dedicated malware and showed how to carry out actual physical destruction (Symantec, 2013) by remotely accessing and altering software. It also revealed the intent and the capacities of certain nation-states to develop and deploy such measures. In recent years states have reacted to this development by developing defensive measures to protect national IT infrastructures, extending national security and military doctrines to provide legal and organizational frameworks and establishing new and dedicated government or military institutions for these tasks. In addition, a large number of countries have also adopted offensive strategies, included those involving cyberspace, in their military planning and have established human and technological capacities (UNIDIR, 2013). This situation was emphasized by similar announcements by different states such as the US (US-DOD, 2018a) and the United Kingdom (UK Government, 2016). In 2016, NATO also declared (NATO, 2016) that incidents involving matters of or in cyberspace could invoke application of Article 5 of the Washington Treaty and prompted its member states to establish necessary military cyber capacities able to defend the alliance in this domain. A further major development was the US adoption of a new *defend forward* cyber security strategy in

2018 (US-DOD, 2018a). Declaring the ineffectiveness of defending the national IT systems by establishing IT security measures for them, the new strategy shifts activities outward to focus on the IT systems of potential adversaries and establishes a persistent engagement of cyber forces. Constant activities within foreign IT systems should, according to the strategy, provide early warning of looming attacks and keep foreign cyber forces busy enough to prevent and deter cyberattacks in the first place (Healey, 2019).

2.1 *The Current Situation of State-Driven Cyberattacks*

When it comes to the application of cyber measures in actual physical warfare, however, it seems that cyberattacks more often play a supporting role in military conflicts and are currently not used for massive destruction but rather for reconnaissance as well as the gathering of combat-relevant information. Most of the known cyber incidents were either cases of espionage, campaigns for political influence (Desouza et al., 2020), targeted minor IT systems or were performed with valid user credentials for critical IT systems gathered via social engineering and classic intelligence work. Although the potential for massive destruction was suspected in some cases, only a few cases with explicitly designed and deployed destructive cyber weapons have been identified so far, such as *Shamoon* (SecureList, 2012) or *TRITON* (Miller et al., 2019), both of which were deployed to sabotage central IT systems of Saudi Arabian petrochemical companies. From a strategic perspective, malicious cyber tools seem to have become widely accepted as an additional measure in hybrid conflicts or similar situations that deliberately stay below the threshold of full-fledged military confrontation. The relatively inexpensive creation of offensive cyber capacities—compared with traditional armament—also empowers new international actors. For instance, the Democratic People’s Republic of Korea (North Korea) has become a relevant actor in cyberspace and has been responsible for different incidents over the last years (Ji-Young et al., 2019) such as the hacking attacks against a subsidiary of Sony, banks in Bangladesh or cryptocurrency marketplaces (US-DHS, 2020). Finally, the trend toward the stockpiling of vulnerabilities and exploits as the *base material* for cyber weapons raises new international threats. Undisclosed vulnerabilities in popular software not only provide possibilities for attacks by the withholding party but, conversely, leave anyone using the product vulnerable to attacks by any actor which becomes aware of the weak spot. The incidents of *WannaCry* (GReAT, 2017) and *NotPetya* (Mimoso, 2017), with their massive damage and commercial losses, are dramatic demonstration of this. Both malware campaigns exploited a vulnerability named *EternalBlue* that had been harbored and stockpiled by the US National Security Agency (Kubovic, 2018). The examples demonstrate on the one hand that states are increasingly developing and deploying offensive cyber capabilities, although trying to avoid serious damage to human life and staying below the threshold of IHL-prohibited aggressive actions. On the other hand, military cyber units are

probably training and preparing for utilization of their capabilities in the event of conflicts. In addition, relatively cheap military cyber capabilities are revealing potential regional power shifts, thus increasing the probability of their application in smaller-scale conflicts.

3 How the Technology of Cyber Weapons and Its Application Will Evolve

A starting point for anticipating the influence and impact of AI/ML on the militarization of cyberspace, is the assessment of the possible evolvement of cyber weapons in general as well as consideration of future challenges regarding this type of technology. With the ever-growing automatization of all kinds of technological processes, IT systems are increasingly being integrated into physical systems and devices to control specific functions. Additionally, these IT systems will be further connected with each other (like the *Internet of Things*) and to cyberspace in order to perform tasks remotely (Russell, 2020). This means that defense against cyberattacks will involve an ever-increasing range of distributed digital devices that need to be made even more resistant against malicious influence, as well as chain effects due to interconnections and dependencies. In addition, with the increasing number of devices and the data they create, process or store, the amount of information that needs to be integrated and processed to detect anomalies and malicious operations will continue to rise. The range of possible attack vectors will further grow and diversify. Given the necessity to react to attacks in (almost) real time, the required decision-making must be accelerated and information processed almost instantly. This requires decision-making based on integrated mechanisms of autonomy or the filtering and pre-processing of information to compensate for the relative slowness and limited capacities of human operators (Burton & Soare, 2019). Moreover, this kind of automatization might possibly lead to a *cyber-vs-cyber* situation, where attacks are directly blocked by dedicated defensive measures without human intervention. Similar early consideration of offensive operations and an automatic infection of possible targets within cyberspace by an NSA-backed program called MONSTERMIND (Zetter, 2014) were exposed by Edward Snowden in 2013. Following the US *defend forward* and *persistent engagement* strategy, which will probably soon be adopted by other states, such developments will result in a further undermining of global IT security by means of the preparatory or precautionary installation of backdoors within foreign IT systems, in order to have the option of deploying the intended payload in time. As cyberspace is, on the one hand, the domain of military activities but, on the other hand, also represents the *physical space* that processes the transmission of any kind of action, the IT infrastructures, being its backbone, will obviously become relevant targets themselves. Finally, as the capability already exists, it is presumably only a matter of time until cyber capacities will be used and deployed openly in fully-fledged military

conflicts, since situations already exist where the IT of military systems and weapons themselves have become targets (Perkovich & Hoffman, 2019).

4 How Artificial Intelligence and Machine Learning Could Influence Cyber Weapons

Reflecting on the possible impact of AI/ML on cyber weapons and the militarization of cyberspace, it is crucial to highlight that cyber and AI/ML measures are *natural siblings*. “[AI and ML] share the idea of using computation as the language for intelligent behaviour” (our italic) (Kersting, 2018). From a purely technological perspective, AI/ML is *just* software: algorithms based on complex computer code that can be integrated into decision processes. Hence, AI/ML is developed and deployed within the same domain as cyber tools and to a considerable extent requires similar know-how in programming, code logic and software life cycle management. In order to be effective, cyber tools must keep pace with the latest technological developments, software updates and the modernization of devices. To reach this level of adaptability and extendibility they are often based on modern development frameworks with modularized, extendable and interchangeable software architecture [see, for example, the *FLAME* malware platform (sKyWIper Analysis Team, 2012)]. Such architecture provides an ideal platform for an extension with AI/ML components. Additionally, computer code offers optimal conditions for creating and facilitating training and testing environments for military AI/ML applications, as the environment can be defined and shaped in every specific detail and according to the intended requirements. This reduces costs and the amount of research and development required. As described in the previous section, an important challenge for cyber as well as other military technologies is the growing amount of information that needs to be processed (Kersting & Meyer, 2018), in contrast to the decreasing time to react to incidents. This dilemma involves incidents within cyberspace but also situations where cyber tools facilitate the analysis of data and the processing of information in order to provide the basis for decision-making concerning physical systems such as weapons or reconnaissance systems. AI/ML algorithms, and especially modern approaches such as deep learning (Charniak, 2019), were developed specifically for cases involving processing large amounts of data, detecting patterns and filtering out relevant information from *digital noise*. According to Schörnig (2018), the “spectrum of possible applications [of AI in the military] ranges from the analysis of trade data to uncover clues for the proliferation of weapons of mass destruction, to the identification of landmines that is boosted by AI with improved ground penetrating radars.” Because of such capabilities, military AI applications are likely to be integrated into cyber tools, as these usually have to deal with a large amount of digital data in trying to detect relevant patterns.

4.1 *Explainability and Responsibility of AI-Enabled Cyber Weapons*

An additional aspect of this development is that the automated conclusion process already mentioned and the resulting selection and decision about actions will be significantly changed when combined with AI/ML algorithms. Whereas the automatization of defensive cyber actions is hardly new, AI/ML are, in the sense of technology which produces an output for a given input without allowing reconstruction of the digital reasoning process or the *line of thought* of the machine or software that led to a specific decision. This creates situations in which the code produces decisions that are no longer deducible and thus prevent humans from intervening based on reasoning. When such AI/ML-enabled measures are used for offensive actions, this creates serious problems in connection with the necessary human integration and interaction (Schwarz, 2019). All these issues have already been the subject of heated debate in connection with autonomous weapon systems (AWS) regarding the responsibility and traceability of decisions (IPRAW, 2019; see the chapter from Anja Dahlmann). In order to address the problem of comprehensible AI/ML decisions, a dedicated field of research (XAI—Explainable Artificial Intelligence) (Gunning et al., 2019) is working on technical concepts that allow human operators either to follow the decisions during the reasoning process (ad-hoc XAI) or the decisions to be recapped once they are made (post-hoc XAI). So far, these approaches are mere theoretical concepts that lack general applicability and are hindered by specific technical features of machine learning such as the distributed and numerical representation of learned information (Barredo Arrieta et al., 2020). Additionally, it is questionable whether ad-hoc explainability can be used meaningfully in an environment characterized by extremely short response times, as the two conditions are mutually exclusive. The speed of reaction in combination with the black-box character of such tools may possibly prevent any opportunity for double-checking of decisions by human operators or for their intervention. Even if the code itself does not *pull the trigger*, human operators might tend to trust the decisions or pre-decisions of machines and follow their suggestions due to a lack of alternatives, time pressure or perceived lack of human influence or oversight (Bajema, 2019). As AI/ML algorithms are trained for specific situations and decisions before they are integrated into productive systems, the operators of the finished application might also be unlikely to know the specific details of the training data, nor have any chance to see, perceive or understand the assumptions and pre-conditions of this data. Besides, this inexplicability could lead to critical junctures in situations marked by high international tension. State actors on the brink of military conflict might lack the ability to communicate and explain automatically triggered actions or conclusions that led to their activities to other conflict parties, thus undermining a valuable measure of immediate conflict reduction. As unlikely as such a scenario currently seems, the discussion of application of AI/ML within the ongoing process of modernization of nuclear weapons arsenals (Field, 2019) is an example that highlights the consequences that are at stake (Boulanin, 2019). The application of AI/ML

for militarized tools within cyberspace reveals an overall similarity to AWS (see the chapter from Anja Dahlmann). The debates on norms and limitations of the application of automated cyber tools could thus benefit from the lessons learned about the human role within the decision-making loop of technological systems and its consequences.

4.2 *AI and the Pitfalls of the Attribution of Cyberattacks*

The black-box character of AI/ML systems could also aggravate other features of cyberspace that are currently considered to be problematic, both in terms of the application of international humanitarian law (IHL) and of established norms of state conduct. One of these features of cyberspace concerns the attribution problem (Rid & Buchanan, 2015). Whereas the possibility of identifying attackers is essential for IHL and the states' right to use military force for self-defense (Grosswald, 2011), this task is complicated, time-consuming, and a forensic challenge due to the technical features of the cyberspace (Riebe et al., 2019). Digital information inherently contains a high degree of ambiguity and virtuality. Information can easily be copied, modified, or actively tailored to set false tracks. Consequently, the meaningfulness of information about cyber incidents needs to be critically evaluated to prevent false assumptions and reactions. Applying AI/ML measures to offensive operations will further reinforce this ambiguity and intensifies the problem of gaining a clear picture of what happened and identifying the actors behind it. The automatic AI/ML-driven evaluation of information about an incident inherently contains the problematic aspect of some conclusions about the origin of an attack being inadvertently misleading and the question of how to react proportionately. Such failure could be triggered either by incorrect or insufficiently trained algorithms, biased input information or by following intentionally created false trails (Herpig, 2019). Although the inner state of an AI is considered a *black box*, this condition is the result of the learning model and the data used to train the AI. Assuming that an attacker obtained knowledge of the model of an applied, static AI/ML and the data which had been used for its training—e.g., through leaks, reconnaissance, hacks, or insecure manufacturers' supply chains—it would be possible to replicate such an AI itself and thus calculate the output that this AI/ML would generate for a specific input. Such knowledge could enable an attacker to tailor its attacks either to avoid detection or to generate incorrect conclusions (Apruzzese et al., 2019). Finally, the development and application of AI/ML in commercial, non-military IT systems, especially in the field of IT security and automated network security surveillance and defense, will produce spill-over effects in military applications. This development will increase acceptance of such systems and put constant pressure on military decision-makers to deploy them to gain a supposed strategic or tactical advantage.

5 The Negative Impact on Arms Control of Artificial Intelligence in Cyber Weapons

The developments outlined above add to the existing challenges involved in applying stabilizing measures in security policy to cyberspace, such as working toward peace-sustaining cyber armament reduction and cyber arms control measures. Firstly, a general problem of cyberspace is its virtual character (Reinhold & Reuter, 2019a). Data has neither a specific geographic location nor a physical representation. It can be reproduced seamlessly and is not limited to a specific and unchanging location but can instead be distributed across different places, such as in cloud applications. As explained above in connection with the problem of data ambiguity, integrating an AI/ML system into existing cyber measures further increases aspects of virtuality and non-tangibility and thus undermines established concepts of arms control even more than software itself already does (Reinhold & Reuter, 2019c). Besides obvious dual-use problems (Riebe & Reuter, 2019), in practical terms the effortless duplication of digital data that concerns ready-made AI/ML applications as well as training data hinders the control of proliferation of military-grade AI/ML technology (see the chapter from Kolja Brockmann). This also negatively affects the ability to measure specific aspects of a regulated item, which is a core requirement of arms control (Burgers & Robinson, 2018). Like cyber tools in general, AI/ML algorithms are computer code, or even more abstractly, structured digital data. They are thus immune to any kind of countability and provide few starting points for measuring parameters that could provide meaningful classification or comparison with permissible thresholds. This missing feature also means a distinction between civil and military AI/ML systems that is capable of going beyond the mere declaration of the intended application cannot be made while also preventing any kind of classification of the capacity and performance of an AI/ML system. This situation constitutes a major obstacle to the development of viable verification approaches for AI/ML applications. Apart from that, as the performance of an AI/ML system depends to a large extent on its training, the question arises as to whether the trade and proliferation regulation of training data—either artificially, as tailor-made datasets or taken from real-life samples and situations—could provide a starting point for arms control and nonproliferation regimes. The chapter “Arms Control for Artificial Intelligence” in this volume on arms control for AI will further evaluate these possibilities and challenges.

6 How Can Artificial Intelligence Support Cyber Arms Control?

Apart from the challenges described above about how AI/ML algorithms can add to the already complicated cyber weapons debates and the attempts at peaceful development in this domain, such technologies could possibly also evolve into useful

tools for cyber arms control and disarmament. In general, AI/ML algorithms are a good tool for combining and processing large amounts of different, heterogeneous, often noisy and rapidly changing data to detect patterns, regularities and *hidden* information (Lück, 2019). A specifically powerful aspect of this technology is the ability to identify similarities within data and find useful matching items that do not fully correspond to the trained items but relate to them with a high degree of certainty. This kind of detection quality is usually a problem that cannot be solved with hard-coded deterministic rules. By contrast, an AI/ML algorithm is able to identify relevant detection parameters during its training phase, establishing a self-developed filter for relevant and irrelevant information. As a result, AI/ML algorithms could prove to be the right tool for managing the information overload of IT systems (Kaufhold et al., 2020) and the challenge of finding the needle in the haystack. Such challenges could be the task of searching for anomalies in information provided by states in the context of confidence-building measures or processing surveillance imagery to detect military installations. A meaningful, currently unexplored application could be to control the proliferation of cyber weapons (Silomon, 2018) by monitoring the distribution and occurrence of specific parts of weaponized computer code. As already mentioned, code can easily be copied and will, in almost all cases, be slightly modified or extended to fit into existing cyber weapons, to work with the specific tools and programming frameworks, or to match specific target criteria. Any detection mechanisms searching for an exact piece of computer code will presumably fail to detect such modified versions. An AI/ML algorithm could be trained to circumvent this problem and to provide at least indicators and probability measures of whether and to what extent computer code matches a specific sample. A similar approach could be used to detect and identify actors behind cyberattacks. Even if this is not directly a task of arms control, it overlaps with the regulation of cyber weapons, because an actor is visible, detectable and identifiable by its behavior, by technical operations performed in foreign IT systems and by the tools employed (Sibi Chakkaravarthy et al., 2019). Whereas it is possible and common to counterfeit these indicators in order to lay a false trail, an AI could be used to detect unconscious similarities of the attackers' style, habits and methods. Institutionalized military cyber actors in particular develop their know-how and the required skills over time. They create, extend and modify their own toolsets and cyber weapon arsenals, which are then reconfigured, combined and adjusted for a specific operation (Olszewski, 2018). This means that specific actors often have *digital fingerprints* regarding their customary tools and hacking strategies. Nearly every cyber activity creates digital traces such as small pieces of code that attackers have previously used to perform their tasks, manipulate files, change system settings or log entries or IP addresses of remote IT systems where data has been copied. Such detectable traces are called *samples* and are already used to compare new code to known samples from prior incidents in order to draw conclusions about an alleged actor. Although captured samples like these rarely match existing samples perfectly, they do contain similarities as they come from the same complex cyber weapon project, use similar methods and approaches, or are more advanced versions of each other. Detecting these similarities and identifying cyber

weapons is a task where AI/ML approaches and algorithms are highly suitable (Roberts, 2019). For example, such identification measures are already used by IT security forensics when analyzing cyber incidents (Kanzig et al., 2019). They are often combined with further indicators such as specific habits and ways of programming, the structuring of computer code or recurring phrases and names. Lastly, the black-box character of AI/ML applications could also be an advantage for arms control measures. An essential element of practical control and compliance monitoring of arms control regimes is the requirement that the actors involved do not want to disclose any sensitive information about the regulated or controlled item (Kütt et al., 2018). This requires technical procedures where participating parties—usually states—are required to disclose as little information as possible when verification is performed and verification devices are developed that conceal all processing steps. In addition, the participating parties would have to be convinced that the results will be reliable and trustworthy. Such a tool, in which a defined input leads to a binary decision of *is or is not a weapon*, could be achieved through AI/ML procedures. To prevent doubts regarding the reliability and the acceptability of the algorithm's decision it would be necessary to prevent any modification or tampering and to preserve the integrity of the algorithm and its trained state. This could be achieved by securing the AI/ML application with *digital seals*, cryptographically calculated unique values—usually very long numbers—like checksums and hashes that represent a specific state of arbitrary digital information. A recalculation of the digital seal would immediately reveal any modification as it would result in a different number if the information has been changed (Putz et al., 2019). These mere outlines of applicable approaches presumably have other peculiarities that need to be taken into account when it comes to real-world applications. Although this issue goes beyond the scope of this chapter, it shows that, despite new challenges, AI/ML approaches can also contribute to arms control.

7 Conclusion

This assessment has provided an overview of the possible development and impact of AI/ML methods on cyber weapons. It is based on current trends and technical AI/ML developments as well as on the already ongoing application of or research on AI/ML in other military fields of operation. The assessment shows that the military application of AI/ML for cyber related tasks will probably exacerbate an already tense situation involving a cyber arms race on the one hand and a lack of international measures to prevent destabilizing and harmful effects on the other. Established measures for arms control, whose application to cyber weapons is already hindered by specific technical features of these tools, will face further challenges. Furthermore, for military decision-makers AI/ML algorithms seem to provide solutions for enhancing their weapon systems and battlefield management capabilities through their ability to integrate, process and refine large amounts of digital data. This could provide a strong incentive for military decision-makers to pursue and apply these

approaches. However, the assessment also showed that, in addition to the necessary questions of peace and conflict research regarding AI/ML in cyber weapons, technological developments reflect ongoing debates about lethal autonomous weapon systems. This makes it possible to participate in these discussions and to benefit from lessons learned. Finally, AI/ML approaches could also provide valuable insights into the challenges of arms control for cyber weapons and help to circumvent some of its technological pitfalls. Either way, artificial intelligence and machine learning are just beginning to find their way into military cyber systems, and the time has come to critically accompany this trend and conduct further research in order to promote peaceful development of cyberspace.

References

- Apruzzese, G., Colajanni, M., Ferretti, L., & Marchetti, M. (2019). *Addressing adversarial attacks against security systems based on machine learning*. In 2019 11th International Conference on Cyber Conflict (CyCon) (pp. 1–18). <https://doi.org/10.23919/CYCON.2019.8756865>
- Bajema, N. E. (2019, November 12). *Can humans resist the allure of machine speed for nuclear weapons?* Retrieved from <https://outrider.org/nuclear-weapons/articles/can-humans-resist-allure-machine-speed-nuclear-weapons/>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Boulanin, V. (2019). *The impact of artificial intelligence on strategic stability and nuclear risk*. Retrieved from <https://www.sipri.org/publications/2019/other-publications/impact-artificial-intelligence-strategic-stability-and-nuclear-risk-volume-i-euro-atlantic>
- Burgers, T., & Robinson, D. R. S. (2018). Keep dreaming: Cyber arms control is not a viable policy option. *Sicherheit und Frieden*, 36(3), 140–145. <https://doi.org/10.5771/0175-274X-2018-3-140>
- Burton, J., & Soare, S. R. (2019). *Understanding the strategic implications of the weaponization of artificial intelligence*. In 2019 11th International Conference on Cyber Conflict (CyCon) (pp. 1–17). <https://doi.org/10.23919/CYCON.2019.8756866>
- Charniak, E. (2019). *Introduction to deep learning*. The MIT Press. <https://dl.acm.org/doi/book/10.5555/3351847>
- Desouza, K. C., Ahmad, A., Naseer, H., & Sharma, M. (2020). Weaponizing information systems for political disruption: The actor, lever, effects, and response taxonomy (ALERT). *Computers and Security*, 88, 101606. <https://doi.org/10.1016/j.cose.2019.101606>
- Field, M. (2019, December 20). *As the US, China, and Russia build new nuclear weapons systems, how will AI be built in?* Retrieved from <https://thebulletin.org/2019/12/as-the-us-china-and-russia-build-new-nuclear-weapons-systems-how-will-ai-be-built-in/>
- GReAT. (2017). WannaCry ransomware used in widespread attacks all over the world. *Securelist. Com*. Retrieved from <https://securelist.com/wannacry-ransomware-used-in-widespread-attacks-all-over-the-world/78351/>
- Grosswald, L. (2011). Cyberattack attribution matters under article 51 of the U.N. Charter. *Brooklyn Journal of International Law*, 36(3), 1151–1181.
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37). <https://doi.org/10.1126/scirobotics.aay7120>
- Healey, J. (2019). The implications of persistent (and permanent) engagement in cyberspace. *Journal of Cybersecurity*, 5(1), 1–25. <https://doi.org/10.1093/cybsec/tyz008>

- Herpig, S. (2019). *Securing artificial intelligence*. Retrieved from https://www.stiftung-nv.de/sites/default/files/securing_artificial_intelligence.pdf
- IPRAW. (2019). *Focus on human control*. Retrieved from https://www.ipraw.org/wp-content/uploads/2019/08/2019-08-09_iPRAW_HumanControl.pdf
- Ji-Young, K., Jong In, L., & Kyoung Gon, K. (2019). *The all-purpose sword: North Korea's cyber operations and strategies*. In 2019 11th International Conference on Cyber Conflict (CyCon) (pp. 1–20). <https://doi.org/10.23919/CYCON.2019.8756954>
- Kanzig, N., Meier, R., Gambazzi, L., Lenders, V., & Vanbever, L. (2019). *Machine learning-based detection of C&C channels with a focus on the locked shields cyber defense exercise*. In 2019 11th International Conference on Cyber Conflict (CyCon) (pp. 1–19). <https://doi.org/10.23919/CYCON.2019.8756814>
- Kaufhold, M.-A., Rupp, N., Reuter, C., & Habdank, M. (2020). Mitigating information overload in social media during conflicts and crises: Design and evaluation of a cross-platform alerting system. *Behaviour and Information Technology*, 39(3), 319–342. <https://doi.org/10.1080/0144929X.2019.1620334>
- Kersting, K. (2018). Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines. *Frontiers in Big Data*, 1, 6. <https://doi.org/10.3389/fdata.2018.00006>
- Kersting, K., & Meyer, U. (2018). From big data to big artificial intelligence? *KI – Künstliche Intelligenz*, 32(1), 3–8. <https://doi.org/10.1007/s13218-017-0523-7>
- Kubovic, O. (2018). *One year later: EternalBlue exploit more popular now than during WannaCryptor outbreak*. ESET. Retrieved from <https://www.welivesecurity.com/2018/05/10/one-year-later-eternalblue-exploit-wannacryptor/>.
- Kütt, M., Götsche, M., & Glaser, A. (2018). Information barrier experimental: Toward a trusted and open-source computing platform for nuclear warhead verification. *Measurement*, 114, 185–190. <https://doi.org/10.1016/j.measurement.2017.09.014>
- Langner, R. (2013). *To kill a centrifuge—A technical analysis of what Stuxnet's creators tried to achieve*. Retrieved from <https://www.langner.com/wp-content/uploads/2017/03/to-kill-a-centrifuge.pdf>.
- Lück, N. (2019). *Machine learning powered artificial intelligence in arms control*. PRIF Report 8/2019. Retrieved from https://www.hsfk.de/fileadmin/HSFK/hsfk_publicationen/prif0819.pdf
- Miller, S., Brubaker, N., Zafra, D. K., & Caban, D. (2019, April 10). *TRITON actor TTP profile, custom attack tools, detections, and ATT&CK mapping*. Retrieved from <https://www.fireeye.com/blog/threat-research/2019/04/triton-actor-ttp-profile-custom-attack-tools-detections.html>
- Mimoso, M. (2017, June 28). New petya distribution vectors bubbling to surface. *Threatpost.Com*. Retrieved from <https://threatpost.com/new-petya-distribution-vectors-bubbling-to-surface/126577/>
- Nakashima, E., & Warrick, J. (2012). Stuxnet was work of U.S. and Israeli experts, officials say. *The Washington Post*. Retrieved from https://www.washingtonpost.com/world/national-security/stuxnet-was-work-of-us-and-israeli-experts-officials-say/2012/06/01/gJQAlnEy6U_story.html
- NATO. (2016). *Warsaw Summit Communiqué: Issued by the Heads of State and Government participating in the meeting of the North Atlantic Council in Warsaw 8-9 July 2016*. Retrieved from http://www.nato.int/cps/en/natohq/official_texts_133169.htm
- Olszewski, B. (2018). Advanced persistent threats as a manifestation of states' military activity in cyber space. *Scientific Journal of the Military University of Land Forces*, 189(3), 57–71. <https://doi.org/10.5604/01.3001.0012.6227>
- Perkovich, G., & Hoffman, W. (2019). From cyber swords to plowshares. In T. de Waal (Ed.), *Think peace: Essays for an age of disorder*. Last retrieved on 03.01.2022, from <https://carnegieeurope.eu/2019/10/14/from-cyber-swords-to-plowshares-pub-80035>
- Putz, B., Menges, F., & Pernul, G. (2019). A secure and auditable logging infrastructure based on a permissioned blockchain. *Computers and Security*, 87, 101602. <https://doi.org/10.1016/j.cose.2019.101602>

- Reinhold, T., & Reuter, C. (2019a). Arms control and its applicability to cyberspace. In C. Reuter (Ed.), *Information technology for peace and security—IT-applications and infrastructures in conflicts, crises, war, and peace* (pp. 207–231). Springer Fachmedien. https://doi.org/10.1007/978-3-658-25652-4_10
- Reinhold, T., & Reuter, C. (2019b). From cyber war to cyber peace. In C. Reuter (Ed.), *Information technology for peace and security—IT-applications and infrastructures in conflicts, crises, war, and peace* (pp. 139–164). Springer Fachmedien. https://doi.org/10.1007/978-3-658-25652-4_7
- Reinhold, T., & Reuter, C. (2019c). Verification in cyberspace. In C. Reuter (Ed.), *Information technology for peace and security—IT-applications and infrastructures in conflicts, crises, war, and peace* (pp. 257–275). Springer Fachmedien. https://doi.org/10.1007/978-3-658-25652-4_12
- Reuter, C. (2019). Information technology for peace and security—IT-applications and infrastructures. In C. Reuter (Ed.), *Information technology for peace and security—IT-applications and infrastructures in conflicts, crises, war, and peace* (pp. 3–9). Springer Fachmedien. https://doi.org/10.1007/978-3-658-25652-4_1
- Rid, T., & Buchanan, B. (2015). Attributing Cyber Attacks. *Journal of Strategic Studies*, 38(1–2), 4–37. <https://doi.org/10.1080/01402390.2014.977382>
- Riebe, T., Kaufhold, M.-A., Kumar, T., Reinhold, T., & Reuter, C. (2019). Threat intelligence application for cyber attribution. In C. Reuter, J. Altmann, M. Götsche, & M. Himmel (Eds.), *Science peace security '19—Proceedings of the interdisciplinary conference on technical peace and security research* (pp. 56–60). TUprints. Retrieved from https://tuprints.ulb.tu-darmstadt.de/9164/2/2019_SciencePeaceSecurity_Proceedings-TUprints.pdf
- Riebe, T., & Reuter, C. (2019). Dual-use and dilemmas for cybersecurity, peace and technology assessment. In C. Reuter (Ed.), *Information technology for peace and security—IT-applications and infrastructures in conflicts, crises, war, and peace* (pp. 165–183). Springer Fachmedien. https://doi.org/10.1007/978-3-658-25652-4_8
- Roberts, P. S. (2019, December 13). AI for peace. *War on the Rocks*. Retrieved from <https://warontherocks.com/2019/12/ai-for-peace/>
- Russell, B. (2020). IoT cyber security. In F. Firouzi, K. Chakrabarty, & S. Nassif (Eds.), *Intelligent internet of things* (pp. 473–512). Springer. https://doi.org/10.1007/978-3-030-30367-9_10
- Schörnig, N. (2018). Artificial intelligence in the military: More than killer robots. In B. Wolff (Ed.), *Whither artificial intelligence? Debating the policy challenges of the upcoming transformation* (pp. 39–44). Science Policy Paper des Mercator Science-Policy Fellowship-Programms.
- Schwarz, E. (2019). Günther Anders in Silicon Valley: Artificial intelligence and moral atrophy. *Thesis Eleven*, 153(1), 94–112. <https://doi.org/10.1177/0725513619863854>
- SecureList. (2012, September 11). *Shamoon the wiper: Further details (part II)*. Retrieved from <https://securelist.com/shamoon-the-wiper-further-details-part-ii/57784/>
- Sibi Chakkaravarthy, S., Sangeetha, D., & Vaidehi, V. (2019). A survey on malware analysis and mitigation techniques. *Computer Science Review*, 32, 1–23. <https://doi.org/10.1016/j.cosrev.2019.01.002>
- Silomon, J. (2018). Software as a weapon: Factors contributing to the development and proliferation. *Journal of Information Warfare*, 17(3), 106–123.
- sKyWiper Analysis Team. (2012). *sKyWiper (a.k.a. Flame a.k.a. Flamer): A complex malware for targeted attacks*. Retrieved from <https://www.crysys.hu/publications/files/skywiper.pdf>
- Symantec. (2013). *Stuxnet 0.5: The missing link*. Retrieved from <https://docs.broadcom.com/doc/stuxnet-missing-link-13-en..>
- UK Government. (2016). *National cyber security strategy 2016–2021*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/567242/national_cyber_security_strategy_2016.pdf
- UNIDIR. (2013). *The cyber index: International security trends and realities*. Retrieved from <https://www.unidir.org/files/publications/pdfs/cyber-index-2013-en-463.pdf>
- US-DHS. (2020). *Guidance on the North Korean cyber threat*. Retrieved from <https://www.us-cert.gov/ncas/alerts/aa20-106a>.

- US-DOD. (2018a). *National cyber strategy*. Last retrieved on 03.01.2022, from <https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/09/National-Cyber-Strategy.pdf>
- US-DOD. (2018b). *Summary of the 2018 Department of Defense AI strategy*. Last retrieved on 03.01.2022, from <https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF>
- Werkner, I.-J., & Schörnig, N. (Eds.). (2019). *Cyberwar – die Digitalisierung der Kriegsführung*. Springer Fachmedien. <https://doi.org/10.1007/978-3-658-27713-0>
- Zetter, K. (2014, August 13). *Meet monstermind, The NSA bot that could wage cyberwar autonomously*. Retrieved from <https://www.wired.com/2014/08/nsa-monstermind-cyberwarfare/>

Drones and Lethal Autonomous Weapon Systems



Anja Dahlmann

● آقای هوش مصنوعی ●

رسانه هوش مصنوعی دانشگاه تهران

@MrArtificialintelligence

Abstract Computational methods such as machine learning and especially artificial intelligence will lend weapon systems a new quality compared with existing ones with automated/autonomous functions. To regulate weapon systems with autonomous functions with the tools of arms control a new approach is necessary: This must be focused on the human role in decision-making processes. Despite this focus, the enabling technologies involve some specific challenges regarding the scope and verification of regulation. While technology will not solve problems created by the use of technology, it may be able to offer some remedies.

1 Introduction

As one type of artificial intelligence (AI), machine learning (ML) exerts its influence on the military realm and arms control in a number of ways. The arguably most controversial application of ML has as its objective inflicting severe physical damage and the killing of humans. This does not imply a self-aware AI-enabled killer machine that deliberately chooses to turn against humans. It is much more likely that it will be a weapon system with weak AI for specific tasks to support the selection of and engagement with targets, leading to attacks without human control by so-called “lethal autonomous weapon systems” (LAWS). Such autonomous functions compress the targeting cycle by eliminating delays resulting from communication and human decision-making during an attack. Consequently, development and implementation of data-driven techniques like AI will make new military options available while posing new challenges for arms control. It will make new decision-making processes to support human capabilities possible, act as a force multiplier, and accelerate the action-reaction cycle on the battlefield. This increase in speed may provide a substantial military advantage. The lack of communications links will also enable operations in secluded or contested environments that would

A. Dahlmann (✉)

Institute for Peace Research and Security Policy, Berlin, Germany

e-mail: dahlmann@ifsh.de

not be possible with remotely piloted systems and would be very dangerous with manned ones. In some circumstances, such as air defense against incoming munitions, target selection and tracking can also be more precise than equivalent actions carried out by a human operator. Such potential military advantages come at the cost of risks of escalation, instability, unpredictability, erosion of international humanitarian law (IHL) and ethical dilemmas.

So far, no common international definition of LAWS¹ exists, but the notion that AWS are armed platforms that can select and attack a target without further human intervention is broadly accepted (ICRC, 2016, p. 1). The targeting functions are directly relevant to the use of force, and hence are referred to as “critical.” Other functions such as mobility, health management, interoperability and battlefield intelligence (Boulanin, 2016, p. 7) are less relevant. This technology-agnostic definition is useful for arms control debates because the enablers of the functions are irrelevant from a legal perspective, because the law addresses the humans and their role during attacks. Nevertheless, certain technologies are necessary for elaborated and flexible fighting capabilities that set future LAWS apart from existing autonomous/automated weapons systems such as counter rocket and artillery and mortar (C-RAM) systems like *Phalanx*, or sentry guns such as *Samsung SGR-AI*, but also in self-destructing drones such as *Harpy*.

The autonomous functions can be implemented in drones, submarines, ships or land-based systems such as tanks and cars. They can be part of single platforms, swarms, human-machine teams or battlefield management systems. Autonomous targeting functions may also become relevant for the use of hypersonic glide vehicles. Due to their high velocity a plasma layer forms around those missiles, rendering direct communication impossible. Autonomous functions might be an option for dealing with this lack of human influence—or to assist in defending against them.

This chapter discusses the role of AI in the development of LAWS, the specific challenges of AI and software for an arms control regulation of LAWS, and situations in which those techniques could be beneficial to arms control.

¹The term “autonomy” can be easily used instead of “automation” without leading to a different outcome. The most precise term for addressing “LAWS” would probably be “automation in the use of force” or “automation in targeting functions”. If not stated otherwise, I use the term “autonomous weapon systems” or “LAWS” as shorthand for this meaning. I apply the term “weapon systems” instead of “weapons” to signify its potential functional and geographic distribution: The various functions do not have to be located on just one platform but can be connected through communication links (Asaro, 2012, p. 690, fn. 7).

2 Drones and LAWS: Technology and Functions

To understand the challenges to arms control posed by LAWS, a look at the enabling technologies of autonomous functions is crucial. Since remotely piloted drones are a predecessor to LAWS and imply a certain logic related to autonomy, they will serve as the starting point for those deliberations.

2.1 *Remotely Piloted Drones: Steps Toward Autonomy*

The term “drone” is usually used for unmanned (or uninhabited) aerial vehicles, regardless of whether they are remotely piloted or have autonomous functions. They are also referred to as “unmanned aerial vehicles” (UAV) or—if armed—“unmanned combat aerial vehicles” (UCAV). While unmanned vehicles can be deployed on the ground, or in or under water, aerial vehicles are the most prominent ones with regard to military use, because the aerospace has fewer physical restrictions and hence is much easier to navigate than other domains.

UAV can be used for surveillance and reconnaissance as well as attacks. Their capabilities have become more advanced in the past few decades. Still, currently deployed UAV are slower than manned aircraft and limited to quite specific scenarios such as surveillance, targeted killings in asymmetric conflicts, border patrols and armed oversight of ground patrols (Fuhrmann & Horowitz, 2017, pp. 402–403). They can, however, fly longer and farther than manned systems, improving situational awareness or range and closing the sensor-to-shooter gap.

The new military options offered by UCAV have sparked the interest of numerous countries: According to Bergen et al., so far 39 countries have procured armed drones, 12 of which have conducted armed drone strikes. Altogether, 29 countries and groups of countries have invested in further research and development of armed drones (Bergen et al., 2020, pp. 2–4). Future developments will most probably enable air-to-air combat with stealth capabilities, improve self-defense capability, and make dynamic manned-unmanned or even unmanned-unmanned teaming possible (for example the demonstrators X-47B, Taranis, nEUROn and the concept for Skyborg). Those capabilities will lead to increasing autonomy in various functions of the weapon system and different steps of the targeting cycle—remotely piloted UCAV are opening up a pathway to more autonomous weapon systems. For example, communication between drone, satellite and operator and vice versa causes a substantial time delay, while the communication link can break down or be hacked (see, e.g., Dickow, 2015, p. 10). By countering such shortcomings an increase in machine autonomy in various functions from navigation to target selection can increase the operational efficiency and effectiveness of UCAV.

2.2 *LAWS: A New Quality of Autonomous Functions Enabled by AI and ML*

The technology behind autonomous functions is a combination of various elements from the field of robotics. A robot is a machine that can sense the environment, process this information, and act accordingly. It carries out these tasks by means of sensors, software, processors and physical means of manipulating the environment and/or moving around. Recent progress in the development and miniaturization of sensors, energy supply, processors and other components as well as software capabilities (e.g., computational methods often referred to as AI and ML to enable more sophisticated pattern recognition) have led to highly improved functionalities of integrated but especially of embedded systems such as drones (see, e.g., Dickow, 2015; Erz, 2020, p. 26). Sensors are, for examples, optical cameras, infrared cameras, hyperspectral and full spectrum imaging (HSI), light detection and ranging (LiDAR), inertial navigation systems (INS) and satellite navigation, and radio detection and ranging (RADAR) (Erz, 2020, pp. 15–22). The data collected by these sensors require different levels of processing power and offer different information such as distance, position or velocity.

The improved capabilities must not, however, be confused with the cognitive capabilities of humans or other living beings. The terms “artificial intelligence” and “machine learning” in particular are anthropomorphizing and misleading and demonstrate the importance of precise terminology in discussing the issue of LAWS. In that regard, roboticist Noel Sharkey warns: “The shadow of mythical AI looms large in the background” (Sharkey, 2012, p. 121). This often leads to an overestimation of capabilities, especially the ability of a weapon system to understand the operational context (Dahlmann & Dickow, 2019, p. 11). Furthermore, the terms “AI” and “ML” lack clear definitions. For example, the European Defense Agency (EDA) had to define a shared least common denominator understanding for its member states regarding defense cooperation: “AI is the capability provided by algorithms of selecting, optimal or sub-optimal choices from a wide possibility space, in order to achieve specific goals by applying different strategies including adaptivity to the surrounding dynamical conditions and learning from own experience, externally supplied or self-generated data” (European Defense Agency, 2020b, p. 36). Ultimately, this definition describes weak AI for solving specific problems.

Two crucial technological capabilities which enable autonomous targeting functions are image recognition (based on, for example, data from electro-optical cameras) and the identification of behavioral patterns: both require ML and often AI to analyze the sensor data. The necessary video footage is provided by camera systems mounted on drones such as the Gorgon Stare program. Only a fraction of the footage can be watched, tagged and analyzed by humans. Here ML and AI can provide further insights. Relevant projects include, for example, Mind’s Eye, Skynet or Maven. They can also be regarded as a preparation for autonomous targeting functions because they (are supposed to) support human decision-making in target selection. *Skynet* is a project by US intelligence agencies to create target lists via

behavioral signatures based on mathematical methods using cellphone metadata. Those signatures supposedly identified potential terrorists who became the target of drone strikes or other sanctions in Pakistan, Yemen and Somalia in 2010 (The Intercept, 2015a, 2015b). In order to do this, the NSA gathered metadata such as call data, user locations or the swap of SIM cards from cellphone users in a certain area and stored them on cloud servers. The National Security Agency (NSA) compared this information with data from known terrorists (training set) in order to identify potential new terrorists. They applied a complex combination of “geospatial, geo-temporal, pattern-of-life, and travel analytics [...] to identify patterns of suspect activity” (The Intercept, 2015a) and used relationships in this data to specify the “likelihood of being a terrorist” (Grothoff & Porup, 2016) using more than 80 different variables. Based on this, the US created a so-called “kill list” (Naughton, 2016). This procedure was not only legally questionable but also technically flawed: The basic assumption that “terrorist behavior” is identifiable and different is doubtful, the method produced many false positives (i.e., people wrongly suspected of being terrorists), the training data set was quite small and the same data was used for training and testing.

Project *Maven* is a collaboration between the Pentagon and private companies such as Google and Palantir. However, following the strong opposition voiced by its employees Google decided not to extend its contract in 2019. Project *Maven* is an image-recognition project using algorithms based on the open-source Google toolbox TensorFlow to analyze masses of drone surveillance footage and convert mere data to intelligence. It is supposed to distinguish between different structures (e.g., detect and classify people, vehicles and other objects) and track them. This helps analysts to gather information and possibly helps prioritize enemy targets. A necessary breakthrough in the development of LAWS would be the creation of a real-time surveillance platform and subsequent battlefield command and control without human involvement (Weisgerber, 2018; Wiggers, 2018).

Data-driven computational methods can effectively support human decision-making in the targeting process. Or, as Saab’s Chief Strategy Officer Christian Hedelin put it: “We come from a world where we throw away most of the data just to find certain signal characteristics, to a future where we will be able to squeeze so much more information out of the data that our sensors gather” (European Defense Agency, 2020a, p. 37). These methods can cause some specific, technical challenges, however: First, the training data needs to be appropriate for the actual use of the algorithms. If the dataset cannot be generalized, is incomplete or not robust, the results will be problematic, such as being biased or unpredictable. While a great deal of civilian/commercial imagery for training algorithms exists and is easily accessible, the data for military purposes is very limited and restricted. This makes it difficult to train AI algorithms for such use cases and can cause the errors described above. For example, self-driving cars struggle with minor (potentially adversarial) alterations to traffic signs and may speed up instead of slowing down or stopping (see, e.g., Xiao et al., 2019, pp. 3968–3969). Because similar issues with adversarial AI could affect military applications, DARPA established for example the project Guaranteeing AI Robustness Against Deception (GARD) to establish theoretical ML

system foundations for identifying system vulnerabilities and creating effective defenses against such attacks (Draper). The unpredictability of AI is addressed by attempts like *Explainable AI* to unpack the black box of neural networks, but it remains a challenge (see, e.g., Voosen, 2017 and for a military context Turek, 2016). Such linear systems are additionally prone to path dependencies that can perpetuate errors throughout the rest of the processing sequence. In the process of target selection those might accumulate and foster unpredictable outcomes (iPRAW, 2017).

To sum up: While AI *can* support human decision-making and significantly enhance operators' situational awareness, it cannot be left up to AI alone to make irreversible decisions such as killing human beings.

3 LAWS Necessitate a New Perspective on Arms Control

To discuss the challenges posed by AI and software for the regulation of LAWS we must consider the current state of the debate on LAWS.

3.1 *The Arms Control Debate on LAWS*

LAWS have been discussed internationally for less than 10 years (even though initial thoughts were published much earlier, see, e.g., Altmann & Gubrud, 2004) in the context of human rights and humanitarian arms control, but the interest of states, NGOs, academia and media has grown rapidly. UN platforms for those debates are the Human Rights Council, the General Assembly and, most importantly, the *Convention on Certain Conventional Weapons*² (CCW). Within the CCW framework and beyond that, various regulatory options exist for addressing (at least some of) the specific challenges posed by LAWS. They range from a legally binding CCW protocol, an international treaty between like-minded states, to soft law measures such as a political declaration or identification of best practices. The different instruments are not exclusive but can be combined, such as by adding best practices for weapon reviews to a CCW protocol.

Most supporters of a ban on LAWS are them from the Global South, Austria is the only EU member state. While many EU member states, along with the EU External Action Service, call for human control in the use of force, they do not endorse hard law measures. Instead, many endorse soft law such as a political declaration. According to its Federal Foreign Office, Germany regards such measures as a first

²Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, concluded in 1980, entered into force 1983, amended in 2001.

step toward a legally binding document (Maas, 2018). The United States and Russia, but also, e.g., Australia, India and Turkey oppose any and all regulation. China stated support for a ban of use but did not follow up on this and defined LAWS as weapon systems without any option whatsoever for intervening. Even though weapon systems lacking such options would be highly problematic, they are also quite unrealistic.

Civil society, most prominently the Campaign to Stop Killer Robots, state representatives and other actors have identified various concerns regarding LAWS from various perspectives. Most of them boil down to lack of situational understanding of the weapon system deployed, of the opaque process, unpredictable outcomes of certain computational methods and the inherent necessity for legal and ethical decisions to be made by humans.

Operational Concerns Major operational concerns of military leaders include fratricide or failing to fulfill operational goals due to lack of situational understanding (iPRAW, 2018a, p. 11). Military decision processes are designed in a way aimed at ensuring the best possible situational awareness during armed conflicts. They can be substantially supported by assisting technologies, but ultimately decisions are taken and responsibility for them accepted by humans (“Legal concerns” below). These specific decision-making processes are often described as targeting cycle, which can be deliberate in the case of a planned action or dynamic when the situation calls for swift action. The application of autonomous functions in this targeting process can present the human soldier with a challenge in making meaningful decisions because of the increased speed required and effects such as automation bias, meaning the tendency of humans to trust computer decisions without any deliberation (iPRAW, 2018a, p. 12; Hughes, 2020; for considerations on the use of LAWS in the NATO targeting process Ekelhof, 2018, especially pp. 21–22). As Hughes argues, instead of “providing near-total situational awareness, AI and automation will make the fog of war much worse for warfighters” (Hughes, 2020).

Legal Concerns With regard to the use of autonomous targeting functions in armed conflicts, several fields of law are pertinent. Primarily, international humanitarian law (IHL) applies. To abide by IHL, LAWS would, for example, have to involve discriminating and proportionate use. While this might be possible in very limited circumstances, many combat situations would probably be too complex and dynamic for a machine to develop an adequate level of situational awareness (see, e.g., Geiß, 2019, pp. 45–46). Closely linked to this is the principle of precaution: during planning and attacks, military decision-makers must adopt all measures necessary to avoid harm to civilians. To be able to adapt to unforeseen situations (that have not been pre-programmed), human judgment will probably be necessary, which calls for a human in the decision loop (Geiß, 2019, pp. 49–50). In this regard, the ICRC also stresses the lack of predictability that comes with autonomous functions (ICRC, 2021, also emphasized by Venkatasubramanian, 2019). In addition to this, the potential delegation of decision-making to machines raises the question whether legally relevant decisions have to be made by a human. The ICRC argues “that the law is addressed to States and humans” and that combatants “will require a

minimum level of human control over weapon systems with autonomy in their critical functions so that they can effectively make legal judgements [...] in specific attacks” (ICRC, 2018b, p. 1; also iPRAW, 2019a, p. 11).

Ethical Concerns The ethical dimension of autonomy in weapon systems has attracted a great deal of philosophical attention because autonomy is a highly-loaded philosophical concept in itself and the transfer of this term to machines is inherently problematic. In addition to this dimension, two other aspects are discussed in the literature: the possibility of designing ethical machines and the role of human dignity in the use of force. Concern about the violation of human dignity is rooted in Kantian ethics. Briefly, killing human beings in war (i.e., violating their right to life) without violating their dignity requires a moral agent who recognizes them as human beings and not just as objects or data points selected by an algorithm (iPRAW, 2018b, p. 12, based on Asaro, 2016). As the ICRC puts it: “The central argument here is that it matters not just if a person is killed and injured but how they are killed and injured” (ICRC, 2018a, p. 10).

Security Concerns Concerns from the security perspective are linked to the operational as well as strategic level. Altmann and Sauer identify two dimensions of instability: First, the “military instability with regard to the proliferation of arms and the emergence of arms races” (Altmann & Sauer, 2017, p. 120), that is, on a regional or global level. Here, proliferation could occur horizontally (from states that have LAWS to those that do not possess the technology) and vertically in “an uncontrolled build-up of arms that drives up military expenditure and exacerbates the security dilemma” (Altmann & Sauer, 2017, pp. 120–121). Second, the operational level includes crisis instability and escalation. Here, the potential increase in operational speed adds to the usual instability in armed conflicts and causes a substantial change in warfare. While the early-warning times for nuclear missiles offer at least minutes to react, the time slot for autonomous weapons could be seconds.

Overall, the use of computational methods in the targeting process raises questions of considerable urgency related to the human role in the use of force. Accordingly, the human-machine interaction becomes a focal point for arms control questions related to LAWS. This constitutes a shift because arms control is usually concerned with the quantity of weapons or military capabilities, not with the targeting process and details of military procedures.

3.2 Human Control: The Decision-Making Process as a New Subject of Regulation

As shown above, the relevant norm for the regulation of LAWS is that of human control over the use of force, which is a technology-agnostic perspective on LAWS.

Nevertheless, the technologies that enable autonomous functions pose a set of specific challenges for arms control regulations which will be discussed below.

Since autonomy in weapon systems relates to software-enabled functions instead of physical platforms, the definition of LAWS is highly disputed. For this reason, a definition that addresses the human role and the decision-making procedures seems to be more sensible than not considering these aspects. Specifying the human involvement is a challenge in itself. Even though the human role might have been an implicit part of the negotiations concerning land mines and weapons of mass destruction (WMD) (Human Rights Watch and Int. Human Rights Clinic, 2016, pp. 10–12), the perceived impending autonomy of machines has trained the spotlight on this issue. The discussion of the human role intensifies—and will possibly replace—the search for a technical definition, but states parties are still, on the one hand, using very similar terms that mean very different things (Ekelhof, 2017), and on the other using different terms as though they were synonymous. Human control in the use of force *can* be understood as a two-step approach enshrined in the design and use of an unmanned weapon system. First, the human operator(s) or commander (s) must gain a situational understanding of the weapon system and the environment before, during, and after an attack. Second, the human must be able to intervene during an attack if any unforeseen changes occur (iPRAW, 2019a, building on Human Rights Watch and Int. Human Rights Clinic, 2012; Roff & Moyes, 2016; Article 36, 2016). The details of implementation depend on the actual context of use, meaning that the level or type of human control adequate for abiding by IHL or ethical standards may vary (Amoroso et al., 2018). While not all states parties adopted this terminology (some actively oppose it, e.g., the United States), the overall concept of human control became a central point of reference in the debates at the CCW meetings.

Consideration of a regulation and its verification in this chapter is based on the assumption that states agree on a norm of human control in the use of force within the framework of humanitarian arms control. ‘Human control’ might be an evolving principle of IHL that could be captured in the preamble to the CCW or the first Additional Protocol to the Geneva Conventions (Rosert, 2017, p. 1). It can be regarded as a new or dormant principle of IHL that should be acknowledged in a legally binding document or politically-binding declaration. More concrete norms can be derived for spelling out details of this rather abstract concept (iPRAW, 2018c, p. 14). To capture this concept, at some point any regulation would have to address military decision-making process (including preceding steps like training), the technology, and the operational context (Campaign to Stop Killer Robots, 2019).

3.3 Specific Challenges Arising from Enabling Technologies

In addition to definitions and new aspects of arms control, regulation of LAWS would have to incorporate the dual-use aspect of the enabling technologies and

would be quite limited in its effectiveness due to properties specific to software and data.

Many components of LAWS will be dual-use, meaning they are used for both civilian and military purposes. In addition, they often originate from the commercial realm. Arms control regulation must not hamper commercial development and peaceful use of AI, as many delegations have stated during CCW meetings. The distinction is probably quite simple with regard to weapon platforms or specific targeting applications and interfaces. The lines blur, however, with regard to infrastructures such as the use of satellites to establish communication links or servers to store data (Erz, 2020, pp. 37–38). From an arms-control perspective, not only commercial use would be acceptable but also military use for functions like navigation or surveillance.

Instruments such as weapon reviews (e.g., Article 36 of the First Additional Protocol to the Geneva Conventions) will have difficulty grasping both software and data-based algorithms. These reviews require states to check that a weapon can be used in accordance with IHL, meaning in at least one use case. Those reviews could be a helpful addition to the norm of human control but would probably be insufficient on their own to deal with LAWS based on data-driven computational methods. For example, the reviews would have problems considering the database and could not deal with systems that learn online (i.e., during use). It is also unclear if every software update that affects the targeting process would be subject to review.

If states agreed on legally binding regulation of LAWS, verification measures could become relevant. The type and effectiveness of such measures is a technological question, but ultimately depends on the political will of states parties. A comprehensive verification of human control over the use of force would be quite challenging and, since differences in the operational context call for different applications of human control, the verification measures would have to be flexible. As initially indicated by Gubrud and Altmann, verification could include measures before, during and after the deployment of a weapon system such as inspections of the design of hardware and software, consideration of rules of engagement or software-enabled tracking of the human role. Quite a few challenges would arise from the characteristics of LAWS, because the software could be altered quite easily after inspection, the human role is a qualitative feature that is hard to define and the autonomous functions are hard to detect from the outside, so that verification tests could not be based on suspicious cases.

A verification regime for LAWS could include various pillars, including a “compliance model based on transparency and confidence-building measures, inspections, technical safeguards, and forensic investigation of suspicious incidents, together with verification of human control and enforcement of accountability in the use of violent force” (Gubrud & Altmann, 2013, p. 4), possibly monitored by a treaty organization. Verifying human control is at the core of this approach. Gubrud and Altmann suggest identifying a set of technical indicators to illustrate the human role during attack. Such indicators could include the existence of suitable hardware to establish a communication link between operator and platform, video feeds from cameras that monitor the operator and software solutions to track actions during an

attack. For reasons of military secrecy, data collected during attacks cannot be made public. Analogous to the black box in airplanes, the data itself could be stored by the nation that deploys the weapon system, while hash codes would provide a digital seal to prove that the data had remained unchanged, which is similar to a blockchain. In suspicious cases, the state in question could provide the actual data recorded (or parts thereof) to a trusted third party that could assess the human role in the attack and check whether the data has been tampered with (Gubrud & Altmann, 2013, pp. 6–7). Even though this approach fits best with remotely piloted drones with increasingly autonomous functions and might have problems with more distributed systems like swarms or battlefield management systems, it is a valuable starting point. This kind of verification measure would “reverse the usual logic of verification (detecting non-compliance) by attempting to continuously track the human involvement in an attack” (iPRAW, 2019b, p. 4).

Computational methods could assist in the analysis of the huge amounts of data resulting from such a monitoring system. Effective verification of human control might only be possible by means of big data evaluation to detect suspicious behavioral patterns or other indicators. If the verification measures included a video feed monitoring the operator (given that this type of human control is suitable for the weapon platform and operational context), AI techniques could, for example, analyze the operator’s eye movements or other behavioral features.

3.4 Positive Impact of AI on Arms Control for LAWS

The assistance of verification measures is just one application where AI and ML could support arms control efforts for LAWS instead of challenging them. Due to the lack of concrete regulation, these deliberations are quite vague and hypothetical. However, real-life examples of similar applications exist: For example, AI and ML have been used to evaluate large databases in order to identify war crimes in Yemen. In connection with that matter, the Organization Yemeni Archive collects open-source data from journalists and citizens as well as social media. It applies blockchain techniques in order to protect the data. Using this database, researchers from the Global Legal Action Network trained algorithms to identify the use of certain types of illegal weapons in this database (Global Legal Action Network, 2020; Hao, 2020). Even though the training requires a great deal of human effort, the subsequent application of the algorithm will be much more efficient than human work alone. This can provide important evidence for human rights organizations in legal trials.

Beyond that, software solutions might be able to install some sort of ethical behavior in machines. At least, that is what Arkin’s concept of an “ethical governor” implies when it translates rules of engagement, which can be based on ethical assumptions, into machine behavior. “This ethical behavioral control approach strives to directly ingrain ethics at the behavioral level, with less reliance on deliberate monitoring to govern overt behavior” (Arkin, 2009, p. 133). The core

algorithm follows rules similar to those the military uses to teach new soldiers. Arkin himself would only deploy this ethical governor in very limited circumstances and environments having a simple structure and admits that the implied assumptions about the situational awareness of the machine are quite optimistic (Arkin, 2009, pp. 126–127).

4 Conclusion

The extensive and increasing use of remotely piloted UCAV is only the first step toward weapon systems with autonomous (targeting) functions in all domains. This development is enabled by developments in various fields of robotics including data-driven computational methods such as AI and ML. This withdrawal of the human yields substantial military advantages but also introduces challenges in operational, legal, ethical and security dimensions. While arms control regulation of LAWS should be technologically agnostic and focused on the human role, it will still be shaped by the specific technological capabilities and limitations of AI. The potential lack of predictability (due to the black-box effect and non-deterministic/probabilistic methods) or situational understanding (due to lack of human cognition, unfit datasets) could be mitigated by an adequate type and level of human control. This should be at the center of any regulation of LAWS.

References

- Altmann, J., & Gubrud, M. (2004). Military, arms control, and security aspects of nanotechnology. In D. Baird, A. Nordmann, & J. Schummer (Eds.), *Discovering the nanoscale* (pp. 269–277). IOS Press.
- Altmann, J., & Sauer, F. (2017). Autonomous weapon systems and strategic stability. *Survival*, 59(5), 117–142.
- Amoroso, D., Sauer, F., Sharkey, N., Suchman, L., & Tamburrini, G. (2018). *Autonomy in weapon systems. The military application of Artificial Intelligence as a litmus test for Germany's new foreign and security policy* (Report Vol. 49). Heinrich Böll Foundation. Retrieved from https://www.boell.de/sites/default/files/boell_autonomy-in-weapon-systems_v04_kommentierbar_1.pdf
- Arkin, R. C. (2009). *Governing lethal behavior in autonomous robots*. CRC Press.
- Article 36. (2016). *Key Elements of Meaningful Human Control. Background paper to comments prepared by Richard Moyes, Managing Partner, Article 36, for the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*. Retrieved from <http://www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf>
- Asaro, P. (2012). On banning autonomous weapon systems. Human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross*, 94(886), 687–709.
- Asaro, P. (2016). Jus Nascendi, robotic weapons and the martens clause. In R. Calo, M. Froomkin, & I. Kerr (Eds.), *Robot law* (pp. 367–386). Edward Elgar Publishing Limited.

- Bergen, P., Salyk-Virk, M., & Sterman, D. (2020). *World of drones*. New America Foundation. Retrieved February 23, 2022 from <https://www.newamerica.org/international-security/reports/world-drones/who-has-what-countries-with-armed-drones>
- Boulanin, V. (2016, December). *Mapping the development of autonomy in weapon systems: A primer on autonomy* (SIPRI Working Paper). Retrieved from <https://www.sipri.org/sites/default/files/Mapping-development-autonomy-in-weapon-systems.pdf>
- Campaign to Stop Killer Robots. (2019, November). Key elements of a treaty on fully autonomous weapons. Retrieved from <https://www.stopkillerrobots.org/wp-content/uploads/2020/03/Key-Elements-of-a-Treaty-on-Fully-Autonomous-Weapons.pdf>
- Dahlmann, A., & Dickow, M. (2019). Preventive regulation of autonomous weapon systems. *Need for action by Germany at various levels* (Research Paper 3). German Institute for International and Security Affairs. Retrieved from https://www.swp-berlin.org/fileadmin/contents/products/research_papers/2019RP03_dnn_dkw.pdf
- Dickow, M. (2015). *Robotik – ein Game Changer für Militär und Sicherheitspolitik?* Deutsches Institut für Internationale Politik und Sicherheit. Retrieved from https://www.swp-berlin.org/fileadmin/contents/products/studien/2015_S14_dkw.pdf
- Draper, B. (n.d.). *Guaranteeing AI robustness against deception (GARD)*. Defense Advanced Research Projects Agency. Retrieved February 23, 2022 <https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception>
- Ekelhof, M. (2017). Complications of a common language. Why it is so hard to talk about autonomous weapons. *Journal of Conflict & Security Law*, 22(2), 311–331.
- Ekelhof, M. (2018). Lifting the fog of targeting: “Autonomous weapons” and human control through the lens of military targeting. *Naval War College Review*, 71(3), 61–94.
- Erz, H. (2020). *Künstliche Intelligenz und Daten. Eine Evaluation softwarebasierter militärischer Informationsgewinnung* (Research Report 4). Institut für Friedensforschung und Sicherheitspolitik.
- European Defense Agency. (2020a). AI has significant potential for autonomous systems. Interview with Saab’s Chief Strategy Officer Christian Hedelin. *European Defense Matters* (19), 37–38. Retrieved from https://eda.europa.eu/docs/default-source/eda-magazine/edm19_web.pdf
- European Defense Agency. (2020b). Joint quest for future defence applications. *European Defense Matters* (19), 34–36. Retrieved from https://eda.europa.eu/docs/default-source/eda-magazine/edm19_web.pdf
- Fuhrmann, M., & Horowitz, M. C. (2017). Droning on: Explaining the proliferation of unmanned aerial vehicles. *International Organization*, 71(2), 397–418. <https://doi.org/10.1017/S0020818317000121>
- Geiß, R. (2019). Autonome Waffensysteme. Ethische und völkerrechtliche Problemstellungen. In I. Werkner & M. Hofheinz (Eds.), *Unbemannte Waffen und ihre ethische Legitimierung* (pp. 41–61). Springer Fachmedien.
- Global Legal Action Network. (2020). *Yemen Airstrike evidence database*. Retrieved February 23, 2022 <https://www.glanlaw.org/airstrike-evidence-database-yemen>
- Grothoff, C., & Porup, J. M. (2016). The NSA’s SKYNET program may be killing thousands of innocent people. *Arstechnica*. Retrieved from <https://arstechnica.com/information-technology/2016/02/the-nsas-sky-net-program-may-be-killing-thousands-of-innocent-people/>
- Gubrud, M., & Altmann, J. (2013). *Compliance measures for an autonomous weapons convention*. *International Committee for Robot Arms Control* (ICRAC working paper). Retrieved from https://icrac.net/wp-content/uploads/2018/04/Gubrud-Altman_Combpliance-Measures-AWC_ICRAC-WP2.pdf
- Hao, K. (2020). Forscher wollen mit Maschinenlernen Kriegsverbrechen dokumentieren und verfolgen. *Heise.de*. Retrieved February 23, 2022 from <https://www.heise.de/hintergrund/Forscher-wollen-mit-Maschinenlernen-Kriegsverbrechen-dokumentieren-und-verfolgen-4835504.html>
- Hughes, Z. (2020). *Fog, friction, and thinking machines*. *War on the rocks*. Retrieved from <https://warontherocks.com/2020/03/fog-friction-and-thinking-machines/>

- Human Rights Watch. (2012). *Losing humanity. The case against killer robots*. Retrieved from https://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf
- Human Rights Watch, & International Human Rights Clinic. (2016). *Killer robots and the concept of meaningful human control*. Memorandum to Convention on Conventional Weapons (CCW) Delegates. Retrieved from https://www.hrw.org/sites/default/files/supporting_resources/robots_meaningful_human_control_final.pdf
- ICRC. (2016). Views of the International Committee of the Red Cross (ICRC) on Autonomous Weapon Systems. Retrieved from <https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system>
- ICRC. (2018a). Ethics and autonomous weapon systems. An ethical basis for human control? Retrieved from <https://www.icrc.org/en/document/ethics-and-autonomous-weapon-systems-ethical-basis-human-control>
- ICRC. (2018b). *Statement of the International Committee of the red Cross (ICRC). Further consideration of the human element in the use of lethal force; aspects of human-machine interaction in the development, deployment and use of emerging technologies in the area of lethal autonomous weapons systems*. Geneva.
- ICRC. (2021). *ICRC position on autonomous weapon systems*. Retrieved from www.icrc.org/en/document/icrc-position-autonomous-weapon-systems
- International Panel on the Regulation of Autonomous Weapons. (2017). *Focus on Computational Methods in the Context of LAWS* (Report No. 2). iPRAW. Retrieved from https://www.ipraw.org/wp-content/uploads/2017/11/2017-11-10_iPRAW_Focus-On-Report-2.pdf
- International Panel on the Regulation of Autonomous Weapons. (2018a). *Focus on the Human Machine Relation in LAWS* (Report No. 3). iPRAW. Retrieved from https://www.ipraw.org/wp-content/uploads/2018/03/2018-03-29_iPRAW_Focus-On-Report-3.pdf
- International Panel on the Regulation of Autonomous Weapons. (2018b). *Focus on Ethical Implications for a Regulation of LAWS* (Report No. 4). iPRAW. Retrieved from https://www.ipraw.org/wp-content/uploads/2018/08/2018-08-17_iPRAW_Focus-On-Report-4.pdf
- International Panel on the Regulation of Autonomous Weapons. (2018c). *Concluding report: Recommendations to the GGE*. iPRAW. Retrieved from https://www.ipraw.org/wp-content/uploads/2018/12/2018-12-14_iPRAW_Concluding-Report.pdf
- International Panel on the Regulation of Autonomous Weapons. (2019a). *Focus on Human Control* (Report No. 5). iPRAW. Retrieved from https://www.ipraw.org/wp-content/uploads/2019/08/2019-08-09_iPRAW_HumanControl.pdf
- International Panel on the Regulation of Autonomous Weapons. (2019b). *Verifying LAWS regulation. Opportunities and challenges* (Working Paper). Berlin: iPRAW. Retrieved from https://www.ipraw.org/wp-content/uploads/2019/08/2019-08-16_iPRAW_Verification.pdf, checked on February 23, 2022.
- Maas, H. (2018). Die Zukunft der nuklearen Ordnung - Herausforderungen für die Diplomatie. *Auswärtiges Amt*. Retrieved from <https://www.auswaertiges-amt.de/de/newsroom/maas-festergarten-konferenz/2112704>
- Naughton, J. (2016, February 21). Death by drone strike, dished out by algorithm. *The Guardian*. Retrieved from <https://www.theguardian.com/commentisfree/2016/feb/21/death-from-above-nia-csa-skynet-algorithm-drones-pakistan>
- Roff, H., & Moyes, R. (2016). *Meaningful human control, artificial intelligence and autonomous weapons: Briefing paper for delegates at the convention on certain conventional weapons meeting of experts on lethal autonomous weapons systems*. Article 36. Retrieved from <http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf>
- Rosert, E. (2017). How to regulate autonomous weapons. Steps to codify meaningful human control as a principle of international humanitarian law. *PRIF Spotlight 6/2017*. Retrieved from https://www.hsfk.de/fileadmin/HSFK/hsfk_publicationen/Spotlight0617.pdf
- Sharkey, N. (2012). Killing made easy. From joysticks and politics. In K. Abney, G. A. Bekey & P. Lin (Eds.), *Robot ethics. The ethical and social implications of robotics* (pp. 111–128). MIT.

- The Intercept. (2015a). *SKYNET: Applying advanced cloud-based behavior analytics*. Retrieved February 23, 2022 <https://theintercept.com/document/2015/05/08/skynet-applying-advanced-cloud-based-behavior-analytics/>
- The Intercept. (2015b). *SKYNET: Courier Detection via Machine Learning*. Retrieved February 23, 2022 from <https://theintercept.com/document/2015/05/08/skynet-courier/>
- Turek, M. (2016). *Explainable Artificial Intelligence (XAI)*. Defense Advanced Research Projects Agency. Retrieved from <https://www.darpa.mil/program/explainable-artificial-intelligence>
- Venkatasubramanian, S. (2019). Structural disconnects between algorithmic decision-making and the law. Humanitarian Law & Policy: ICRC Blog. Retrieved from <https://blogs.icrc.org/law-and-policy/2019/04/25/structural-disconnects-algorithmic-decision-making-law/>
- Voosen, P. (2017). How AI detectives are cracking open the black box of deep learning. *Science*. Retrieved from <https://www.sciencemag.org/news/2017/07/how-ai-detectives-are-cracking-open-black-box-deep-learning>
- Weisgerber, M. (2018). General: Project Maven is just the beginning of the Military's use of AI. *Defense One*. Retrieved from <https://www.defenseone.com/technology/2018/06/general-project-maven-just-beginning-militarys-use-ai/149363/>
- Wiggers, K. (2018, May 29). The pentagon wants to expand its controversial project Maven AI initiative. *VentureBeat*. Retrieved from <https://venturebeat.com/2018/05/29/the-pentagon-wants-to-expand-its-controversial-project-maven-ai-initiative/>
- Xiao, C., Deng, R., Li, B., Lee, T., Edwards, B., Yi, J., Molloy, I., et al. (2019). AdvIT: Adversarial frames identifier based on temporal consistency in videos. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3967–3976). <https://doi.org/10.1109/ICCV.2019.00407>.

No, Not That Verification: Challenges Posed by Testing, Evaluation, Validation and Verification of Artificial Intelligence in Weapon Systems



● آقای هوش مصنوعی ●

رسانه هوش مصنوعی دانشگاه تهران

Maaïke Verbruggen 

@MrArtificialintelligence

Abstract Over the past decade, research in Artificial Intelligence has advanced significantly, but many challenges still remain. One underexplored problem is the fact that it's extremely difficult to test AI. In fact, there are no techniques in existence that can validate and verify AI, to ensure that the systems will function as specified and avoid unpredictable behaviour. Additionally, most military innovation and procurement protocols are designed with hardware in mind, not software, like most contemporary AI. To elucidate this problem, this chapter will set out the challenges related to testing AI, and the implications for arms control.

1 Introduction

When it comes to Artificial Intelligence (AI), it is hard to say which people talk about more: how much it will revolutionize everything, or how great the hazards are. The literature contains many discussions of bias, fairness, safety, security, understandability, explainability and brittleness—the list goes on and on. Silicon Valley is known for its unofficial “disrupt and break things” motto, or its tendency to roll out products first and ask questions later. But defense organizations generally follow a completely opposite approach, and military products go through extensive cycles of Testing, Evaluation, Validation and Verification (TEV&V). Particularly important is the Validation and Verification (V&V) process, which formally proves that a weapon works properly and will function as specified. In an ideal world this could serve as a quality check to ensure that the aforementioned problems with AI such as bias and brittleness would not plague weapon systems too.

Unfortunately, that is not the case. It is not (yet) possible to formally verify and validate AI due to its non-deterministic nature. Moreover, its integration in weapons can have unexpected effects at the system level, called emergence. Consequently, the behavior of a weapon with AI might not be fully predictable. It is not only

M. Verbruggen (✉)
Vrije Universiteit Brussel, Brussels, Belgium
e-mail: Maaïke.Verbruggen@vub.be

impossible to be certain how AI will function, but we cannot even assess how (un)certain we are. Because AI responds to its environment, extensive and iterative operational testing is required to assess how it will function in different scenarios. However, operational testing is generally conducted long after the design of the weapon has been determined and the first prototypes have been rolled out, making it hard to change anything structurally. This does not mean that AI is fundamentally unreliable or unpredictable, but that we struggle to *know* AI and be *certain* of how (un)reliable and (un)predictable it will be.

Ironically, if human control over a system is reduced or removed, or if a system is piloted remotely, it becomes absolutely essential to be 100% sure specifications have been met, because humans are no longer there as a failsafe. Without such a safety net, it is necessary not only to verify and validate the properties of the weapon in a technical sense, also to verify and validate that it will achieve mission requirements. The conundrum is thus that V&V of AI is both more difficult and more important than ever before.

This chapter will focus on Research and Development (R&D), TEV&V and acquisition processes in defense organizations at large, but of course there are stark differences among countries. The vast majority of the literature is from the US, which unfortunately means the way TEV&V is presented here is biased. But even though the literature is US-dominated, the problems with V&V of AI have truly global proportions. Not only does it pose great risks for accidents in warfare, for which civilians will pay the price, as they always do. It also leads to political contestation over what it means to validate, verify and certify autonomous technologies, as countries black box their AI when exporting weapons. This will make it very difficult for arms importers to independently validate, verify and certify the weapons they buy; determine whether these weapons will reflect their national doctrines concerning the use of AI; or conduct Article 36 reviews to assess their legality.

This chapter will start with an explanation of the TEV&V process, and why this is so critical for military technologies. It will then explain the problems with V&V of AI, and how it does not fit existing TEV&V protocols. Finally, a discussion of implications for arms control and potential solutions concludes the chapter.

2 TEV&V: The Traditional Way

2.1 Military TVVE

V&V does not come cheap, lengthens the R&D timeframe, and requires specialized expertise (Gutmann, 2004). Nevertheless, nobody really tests like the military. Military innovation is often highly ambitious, aiming to develop a future weapon with yet to be discovered, mature or proven technologies. Development itself can take up to 30 years, and the weapons are expected to last up to another 50 years. Weapon design thus involves incorporating technologies you only expect to exist in

15 years, and building a system that will remain functional for decades in a hostile environment. In addition, the stakes are very high. If a country's defenses against an invasion fail, it is done for. If your gun jams on the battlefield, you are dead. Losing soldiers can have serious political costs, especially in countries with a casualty-averse culture or ones that glorify the military. Ironically, this intolerance for risks both heightens the importance of V&V and is one of the drivers towards AI and unmanned systems.

2.2 *Verification and Validation*

Verification procedures test whether a system meets design specifications, while validation procedures test whether a system meets the requirements of the user. In simple terms, verification answers the question “*Did you build system correctly?*” while validation answers the question “*Did you build the correct system?*” (Clark, 2015). This is different from arms control style verification, which refers to ensuring that countries comply with arms control.

The benefit of V&V is the depth and extent of certainty it provides. Normal program testing cannot be exhaustive, because it can reveal bugs, cannot show their absence (Dijkstra, 1972). V&V is far from flawless and does not lead to products that work flawlessly. Instead it ensures design and testing is conducted more holistically and reduces the chances of mistakes during the R&D process. Moreover, you get better insight into how your weapon functions. If you know that particular components are prone to failure or hard to verify, you can build in mitigation measures for the problematic parts. If you know the error rate, you can better assess whether a specific use is justifiable, and you can exclude risky use-cases (Luckcuck et al., 2019). What follows is a brief explanation of the role of V&V during the R&D process.

It is essential to first gather the requirements from the users and/or buyers in a structural manner. A distinction can be made between functional requirements (e.g., shooting a rocket beyond the line-of-sight) and non-functional requirements that describe system properties and constraints (e.g., reliability or safety) (Pereira & Thomas, 2020). This stage should also involve a legal or Article 36 review, which is mandatory for states party to the First Additional Protocol of the Geneva Conventions¹ (Boulanin & Verbruggen, 2017). These requirements are turned into specifications, which transform abstract concepts in natural languages into exact specifications, often in technical language (Wayne, 2019). This includes the type of metrics to use and the benchmarks that the system should meet. For example, the user might insist the system should be safe, but “safety” is not something you can

¹While the USA has not signed or ratified this protocol, it does mandate legal reviews for all new weapons.

confirm. It should thus be expressed in a more concrete metric, such as the rate of misfires, and include a threshold, such as once out of 100 times.

Because the production of military systems is extremely expensive, it is very important to be sure the product works before you start manufacturing it in large quantities. Therefore, during operational TEV&V, the product is tested to explore all the potential errors and failures both in use and when interacting with the environment (Keane & Joiner, 2020).

Once the users or buyers are satisfied, the product should undergo another Article 36 review and then become certified. The exact certification protocols differ from country to country. If the product is sold abroad, the end-user will also need to conduct TEV&V. If Life Fire TEV&V is unfeasible or too expensive, formal V&V is critical to assessing whether the product works correctly, and whether it satisfies the specified needs (Keane & Joiner, 2020).

2.3 Methods

V&V can be conducted using a variety of methods, which can be placed on a spectrum of correctness. This means that there is no single specific way V&V must be conducted. The most reliable and correct methods are *theorem proving* and *model checking*. Together they are called *formal methods*. Formal methods lead to safer systems and provide a “legal burden of proof regarding due diligence” (Schaffer & Voas, 2016). They provide proof that the system meets the specifications, and assess the consistency, completeness and correctness of a system in a systematic fashion (Gutmann, 2004).

Theorem proving is a highly abstract process in which mathematicians formally prove that a system will display the properties the model predicts, based on the external circumstances and the previous states of the system. Model checking involves running a digital model of the system through all different possible states to assess whether it exhibits the specified behavior. The goal is to search for violations of this behavior, in order to remove bugs and ensure the system runs consistently. Model checking is less “correct” than theorem proving because this information is not as “total” as mathematical proof, but it is easier to use, and it is easier to identify the specific bugs causing problems.

Another tool is *runtime verification*. All events, states and variable updates of the system are compared with the formal specification that sets out expected behavior. If there are discrepancies, the machine can take action to correct them, notify the operator or, in the case of an adaptive system, fall back onto a deterministic and predefined control system. This is thus very useful for operational TEV&V. However, it takes extra memory, CPU and bandwidth, which makes its use in smaller weapon systems such as tactical drones problematic (Bhattacharyya et al., 2015).

The past three decades have also seen tremendous improvements in modeling and simulation (M&S) capabilities, spurred by advances in high performance computing

as a replacement for nuclear-weapons testing. M&S makes it possible to experiment with different weapon designs, conduct early user testing, vary the environments of operational testing, and expose a system to particularly risky scenarios to see how it holds up. However, these models must also be verified and validated themselves. Verifying digital representations of the world is not so easy . . . which brings us back to the topic of AI.

3 What Is AI?

This paper employs a broad definition of AI, owing to the diversity of the field. Over its 70 year history, AI has been pursued from many different angles, both diachronically and synchronically. Drawing a line that distinguishes AI, automation and software is thus hard, and the scope of the various terms has changed repeatedly over time. Successful applications often become normalized over time and cease to be considered AI, such as Automatic Target Recognition systems or search functionality. The popularity of Machine Learning (ML) has brought the subject back into the public spotlight, but the problems with TEV&V of AI are much broader and far older. However, ML and particularly Deep Learning (DL) have introduced a whole additional set of issues, as though things were not already complicated enough.

The problems with TEV&V of AI affect a broad range of applications. We can identify four general but overlapping functions of AI in weapon systems. First is the function of processing input data. This includes tasks such as object identification, 3D mapping, data filtering and sensor fusion. After being processed, data can be redirected to other components, other systems or to the operator. Second is the control function: monitoring the internal and external environment to maintain stability of the system. This includes self-repair, signaling the need for predictive maintenance to the warehouse crew, or sense-and-avoid algorithms where the system responds to the environment. The third function is the task of planning, which refers to *how* a set goal should be achieved. For example, if a robot is assigned to go from A to B, it can plan a route itself. A human-controlled system can include recommendations on how to take action in order to achieve a preassigned goal. The fourth function is executive decision-making. There is no strict boundary between planning and decision-making, because any distinction is based on the level of specificity of the assigned goals. If a missile is assigned to hit a very specific object, or specific geographic coordinates or a specific electromagnetic signature, but autonomously decides on the angle of impact, the exact moment of explosion (e.g., just before or just after impact, based on the target material), or the flight path in the case of a moving target, it might make more sense to speak of “planning.” If the missile is assigned very broad parameters (e.g., enemy tanks in area x), and the missile selects the exact target, it might make more sense to speak of “decision-making.”

Obviously, this boundary is extremely blurred, which makes the debates on Lethal Autonomous Weapon Systems (LAWS) complicated. AI has been used for

both planning (e.g., fire-and-forget missiles) and decision-making (e.g., air defense systems) in weapons for at least 70 years. But as AI has improved, the range of freedom within which the system is allowed to operate has widened. The question is where and how to draw the line. The struggles with TEV&V of AI further complicate this question. But this does not just apply to LAWS, as it makes all military systems with AI less knowable and predictable..

The following section will set out the problems intrinsic to AI, ML/DL and system integration. It is followed by an explanation of why AI is so difficult to test, especially with our current protocols. In conclusion, the chapter describes the implications for arms control.

4 Problems with the Technology

4.1 *The Logics of AI*

The key problem of TEV&V of AI is that many techniques operate in a non-deterministic fashion. The exact regulations differ from country to country, but US laws specify that all safety-critical software in military aviation must be deterministic and time invariant. This means that the exact same configuration of input must always lead to the same output and cannot change over time (Lyons et al., 2017). Obviously, all weapon components will behave differently based on the environmental conditions. But adaptive systems vary not only in their output based on the input, but also in the way the input is transformed into output by utilizing non-deterministic logic. For example, swarm-control algorithms determine the starting position of the units of a swarm stochastically (randomly), because otherwise every node would move to the exact same spot. Many AI algorithms are intrinsically probabilistic in their programming, execution or data sampling, and all neural networks produce only approximate solutions to the problems they have to address. ML systems that learn online (after deployment) will function differently over time, and are thus not time invariant (Braiek & Khomh, 2020).

Because the logic is indeterminate, a system can find itself in one of a near infinite number of different states. This is called the *state space explosion*. A state refers to the values assigned to a program variable, and the state space refers to the set of all different possible states of a program. Formal methods cannot exhaustively search, examine and/or test all different states. It is possible to search through subsections of the state space, but doing so risks missing critical interaction effects between subsections (Haugh et al., 2018). Moreover, it is also not necessarily clear how these subsections causally relate to each other.

And while all systems respond to the environment, most do not respond so sensitively to humans as some AI systems do. In fact, one of the most important environmental factors is the *human-machine interaction*. Humans and machines interact with and respond to each other. But humans are highly unpredictable, which makes it hard to foresee all different ways a machine might respond (Bolton

et al., 2013). This is especially pertinent in manned-unmanned teaming set-ups, because these rely on a shared understanding between human and machine about mission goals and methods. But any mathematical model that tries to capture the way humans reason will inevitably fail, which means that we cannot mathematically verify how machines would act in all possible scenarios. It is important to realize that the problems with V&V are not dependent on the level of autonomy of a system. In fact, systems that have intermediate levels of autonomy and are used in close interaction with humans are the hardest to validate and verify due to the unpredictability of the human element (Tate, 2019b).

4.2 *Developing ML*

The leading concept of the third AI summer in the 2010s–2020s is a particularly gnarly type of AI called ML. ML is developed by applying a model with a general framework and specific computational techniques (such as regression or Bayesian networks) to a (large) dataset. The model searches for patterns in the structural features of the data, as opposed to its semantic content. The structural features in the data that explain a pattern are then reduced to the most essential characteristics and can then be applied to other data, where the model will find these same patterns. This constitutes the ML algorithm that you will integrate into your weapon system.

The core problem for V&V is the question whether the pattern of structural inferences you have found actually also appears in the same way elsewhere. This is the question of the *generalizability* of ML, which is extremely critical, but there are no metrics for measuring it (Mahajan et al., 2020). Thus, ML might detect patterns where there are none. Paradoxically, an ML algorithm is often also hyper-deterministic at the micro-level (e.g., which pixels enable classification). That means that small changes in how the data is structured—that would be irrelevant for human interpretations—can lead to wildly different conclusions by the machine (Wojton et al., 2020). Even different ways of formatting data can drastically reduce the performance of an algorithm. Mahajan et al. (2020) show how a ML algorithm trained to detect knee ligament tears trained on MRIs from one hospital, dropped 15 percentage points in performance when applied to MRIs from another hospital. Meanwhile, weapon systems with AI are supposed to function not only in highly heterogeneous networks of different weapon systems, but in transnational military operations with weapon systems of different countries. This is what makes it so critical for weapon systems to be interoperable, and for different data sets to be preprocessed and calibrated. It makes ML in its current state extremely brittle, as the chance of failure if it is used in conditions other than the conditions specified is very high. However, validating and verifying this condition—called *robustness*—requires knowing under what condition the algorithm will and will not perform. This means that not only the algorithm must be validated and verified, but also the data used to train it. It is necessary to confirm that the data is structurally similar to and representative of the data encountered in real life (Pereira & Thomas, 2020).

Data that is not representative is called *biased*. In practice the training data and the real-world data are almost never identically distributed, which is called the *reality gap* (Luckcuck et al., 2019). Even if it would be possible to verify and validate this at the moment of deployment, many ML algorithms continue to learn *online*, while in service. Not only does this mean that the algorithm will continuously change, but we also cannot validate or verify the extent to which it will change. We also do not know what sorts of data it might encounter after deployment, and to what extent this will differ from the training data. This process of continuing to learn and thus moving away from the requirements is called *model drift* (Pereira & Thomas, 2020). Model drift makes the algorithm inherently unpredictable, can lead to decreased performance or even *catastrophic forgetting*, without the operators noticing (Wojton et al., 2020). While online learning makes the ML less predictable, it can increase its *accuracy*. Real-life data could be more representative than the training data, and ML algorithms become more accurate the more data they train on (Pereira & Thomas, 2020). The predictability that V&V requires must thus be weighed against accuracy, which is no easy choice.

Not only are we uncertain about the performance of ML, we are even uncertain about how uncertain we are. Training ML algorithms always involves setting initial parameters in the model, and tuning these hyperparameters between training sessions. These parameters demarcate a subset of the model space to search through and focus on for training (Braiek & Khomh, 2020). These parameters are generally based on the suggestions from the framework, manual configuration based on experience or the domain literature, or literal trial and error. But the exact relation between the (hyper)parameters and the performance of the algorithm is not really understood. Consequently, traceability is impossible because of the unclear connection between the algorithm, the model, and the data. Moreover, the standard technique in software for assessing how extensively you have tested (to compensate for the lack of completeness if formal methods are not being used) is called *coverage testing* (Braiek & Khomh, 2020). Test coverage describes the extent to which the tests have confirmed all the requirements, and code coverage describes how much of the code has been run in the conduct of tests, to find those rarely used pieces of code and assure they also function properly. *Code review* by bots or preferably by external people helps remove errors and bugs. All of this is not possible due to the indirect nature of programming ML algorithms and the near infinite state space that is to be explored (Varshney & Alemzadeh, 2017). We are thus not only unable to verify that the weapon matches its specifications but also cannot compensate by formalizing and streamlining the work process, as is done in, for example, aerospace engineering through the extensive use of checklists.

The situation is even worse for DL. Most ML models require the programmer to dictate in advance which features of the data the algorithm should mine in the search for patterns, an activity called *feature engineering* (Braiek & Khomh, 2020). This contrasts with DL, where these features are inferred by the model itself. This makes it possible to uncover patterns based on logics that are extremely difficult to codify. The advantage of DL is that you can develop AI for those functions that come naturally to humans, and are consequently difficult to codify. For example, people

know how to speak, read, or identify faces, but if anyone asked them how they do it they would be at a loss. Because the point of DL is to develop algorithms for those capabilities we struggle to program, setting out specifically how the algorithm should function so that we can validate and verify it would defeat the entire purpose even if we knew how to do it. And we do not know how to do it, because the DL algorithm reasons in such an alien way that humans struggle to grasp it. This opacity is what is often called the “black box” of AI, and it has two aspects: explainability and interpretability. Explainability describes how well we can understand the way in which a machine comes to a conclusion. For example, often we do not know which features a machine uses to classify an object. All we have is a set of weighted integers in the algorithm which are difficult to grasp for humans. This is particularly a problem for the user. Interpretability describes how well we understand why the machine comes to their conclusion. For example, we often do not understand *why* a machine is using those features to classify a picture. We struggle to identify the causal inference between the settings of our training model, the data and the exact nature of the algorithm this results in (Braiek & Khomh, 2020). Not only must we grapple with not understanding how the DL reasons, which makes it impossible to model, validate and verify, but even if the algorithm shows high levels of accuracy on the training data, we do not know whether the system came to the right conclusion for the right reasons. For example, it might not fire because it did not detect anything at all instead of classifying an object as a hospital. But we need to understand *why* a system makes a certain decision to verify and validate its generalizability on anything else (Wojton et al., 2020).

4.3 Integrating AI

Ideally, all software involving AI would be modeled in an integrated manner in advance when designing a new complex weapon system. This might be the way weapons are developed in the future. But for now much of current development focuses on developing individual computer programs or retrofitting existing systems to add autonomous capabilities (Reim, 2019). AI is often seen as a component that can be integrated into a system at a later stage, independent of sensors, processing power, communication equipment, etc. But these components can affect the computational intelligence of the AI in question. For example, most ML algorithms are trained using extremely powerful parallel graphic processing units. However, it is extremely unlikely that the weapon system the algorithm will eventually be integrated into will be a system running on parallel processing of equal power. While running the algorithm is not as computationally intensive as training the algorithm, it is still necessary to ensure that the system will have sufficient capability to run properly, as this might otherwise result in unexpected time lag (Pereira & Thomas, 2020). Another example is the situation where the algorithm analyzes data from different sensors in the weapon system, but this requires proper calibration (Young, 2016).

In addition, AI has a penchant for generating emergent properties. There are many different definitions of emergence, but here we will talk about *weak* emergence, and define it as unexpected properties or behavior at the system level. All complex systems are prone to these weak emergent properties (Johnson, 2006). From a design perspective, emergence can result from unexpected interactions between components or between components and the environment or from a product design that insufficiently specifies the properties of subsystems and components or interfaces between them, leading to unexpected combinations. From an organizational perspective, emergence is generally the result of subcontracting work on different components and subsystems. This inhibits gaining holistic understanding of the system and obstructs synchronization and coordination between different subsystem and component development processes and actors (Brusoni & Prencipe, 2011).

Emergent behavior has become an increasingly common issue with the introduction of software into weapon systems (Clark, 2015). This radically increases the number of possible interactions and linkages that must be validated and verified, and this more complex interdependence creates many new opportunities for failure (Hylving & Schultze, 2020). In addition, weapon systems are increasingly coupled in extensive networks of systems of systems linking the entire battlefield together. As shown above, AI is generally unpredictable as a component because its behavior depends on the state of the environment, which cannot be fully known in advance. Moreover, it links different components, and consequently radically increases the complexity of the system as a whole. As such, US Army representatives have stated that identification and evaluation of emergent behavior is more important than finding defects (Deonandan et al., 2010).

The introduction of software and the rise of complex systems has already created many difficulties for TVVE that have not been fully solved. And when problems appear, it can be difficult to isolate their cause. Small changes in the design of the components can cause significant changes in system behavior, but it is often unclear exactly which factors caused the changes (Tate, 2019a, b). This can be grasped well using a sports team as an analogy: You can easily tell whether a team is doing well or not, but exactly what enables or hinders success is much harder to identify, let alone correct.² Most of the time, emergence is an unknown unknown. Theorem proving is particularly inadequate here—but it cannot check something that was not part of the design. Model checking is a better option, but it still becomes very difficult to check all the different states of complex systems at the system level (Bolton et al., 2013). Consequently, most V&V of complex weapon systems concentrates on formal verification of a small number of specific data points that have been identified as most critical, based on the underlying theoretical model. But since we do not know fully how AI reasons, we do not know which data points are most critical or whether they are potentially coupled with other data points.

²Except of course that it's always the fault of the referee.

5 Problems with the Process

5.1 Testing AI

Most acquisition protocols are designed in a sequential fashion. The exact stages differ per country and weapon, but generally speaking, acquisition involves separate phases such as modelling the requirements, designing the overall architecture of the weapon system, decomposing the design into sub-systems, and continuing this process on a lower level. Every phase is validated and verified and each subsequent phase fills in more detail and validates and verifies it again. The problem is that this protocol has been designed with hardware in mind. It is done to reduce the chance of errors further down the process, where they get exponentially more expensive—particularly if they are found in the production phase, as this is the most expensive phase of R&D. By contrast, software has no production phase, as it does not need to be manufactured. Obviously, it will still need to be implemented somewhere, but that is equally true for hardware, and implementation is different from production. As such, the lifecycle of R&D in software is different than R&D in hardware. However, most acquisition protocols have been designed with hardware in mind.

The timelines of R&D of AI and traditional weapon systems vary greatly. Major complex weapon systems can take up to 30 years to develop and are meant to last for several decades. AI is developed much more rapidly, is expected to be in use for a far shorter time and its state-of-the-art advances more rapidly. This poses a substantial problem for defense planning. The exact shape and content of models always evolve over the course of the R&D trajectory, but the skeletal structures are set in stone, because of the high investment costs necessary to develop the platform. Systems currently nearing deployment were not necessarily built with AI in mind. Moreover, the countries now commencing R&D on a system will have to spell out requirements for what AI should be capable of in 20 years, which is quite a gamble considering both the cyclical long-term and fast-paced short-term history of AI. In addition, it means that the TVVE test designs and benchmarks will be known years in advance. Implicitly and/or explicitly, developers and AI might use this time to train the algorithm to pass the test. Stories of so-called “reward hacking” occur throughout the history of AI, and they are often only uncovered much later (Wojton et al., 2020).

Because it is difficult to even conduct TEV&V on an isolated piece of software with AI, many problems only become apparent after system integration. However, this is quite late in the R&D process, after the specifications, product design and other components have already been developed. This makes it both very costly, and more difficult to correct errors (Cook & Haverkamp, 2020). Tallant et al. (2006) estimated in 2006 that 60% of all development costs were spent on control law, software implementation and tests; and that the rise of intelligent and adaptive control systems will increase 200% for single-vehicle and 300% for multi-vehicle control systems. Clark (2015) estimates that 50% of faults are found in the integration and system-test phase, but these errors are 20 times more expensive to correct than if they had been detected in the design phase. He also estimates that 70% of

software development costs are for rework and certification. This has three consequences. Not only is it technologically harder to fix these errors but it is rare that sufficient time and money has been budgeted to do it properly, and the costs of the project will already be so high that defense organizations will be less likely to return to the drawing board to fix the problems in a systematic fashion.

This problem is particularly bad for ML, because the code behind the algorithm is not at all structured and decomposable. This means that “if you change anything, you change everything” (Sculley et al., 2015). There are no individual parts of the code to correct, and you can only investigate the ML algorithm as it exists (Pereira & Thomas, 2020). If bugs are found, it is much more difficult to trace them back to the source (Braiek & Khomh, 2020). And if bug fixes are applied, the entire ML algorithm has to be tested all over again, as a small change can completely alter the way the ML algorithm functions.

5.2 *Measuring AI*

Another problem is that we are not really sure how or what to measure. V&V requires a baseline truth to assess a model’s veracity. For material systems, this is a mixture of theoretical knowledge and historical data. For example, due to the standardization of components, it is known that bolt X has an average failure rate of Y. This is accounted for and redundancies and resilience is built into the system to make it more robust. While software packages are common, it is rare that entire subsections of a program are reused, so there is no historical data to build upon. Moreover, there is no equivalent theoretical base for AI. In fact, when it comes to DL, we actually do not understand why it works so well. This gives us no baseline to compare with. Even at a regulatory level there are no publicly known defense TVVE frameworks for AI, with protocols, benchmarks or standards (Wojton et al., 2020).

Additionally, a whole new type of characteristic must be validated and verified. Civilian certification of products is focused on the technical capabilities to ensure safety, while military certification involves both technical and performance capabilities (Langella, 2013). But the introduction of AI also mandates certifying operational capabilities in order to see how well a system can execute a mission. And there is rarely a baseline truth or a “right” way to execute a mission. Consequently, AI lacks a *test oracle* to test against to ensure its correctness (Bhattacharyya et al., 2015).

The ML community is absolutely no help here, because proper metrics for assessing the functionality of ML are sorely lacking. The performance metrics most commonly used are accuracy, recall and precision. However, all of these refer only to how accurate the model is compared with the testing data, and presume equal distributions among classes. But ML as a discipline rarely even utilizes the more sophisticated statistical metrics that already exist. Metrics for basic functions such as generalizability do not exist, nor do we know how to assess human-machine teaming (Handelman et al., 2019). It is also true that there are no proper benchmarks

for assessing how DL performs based on the hardware-software configuration, which would be essential for V&V of cyber-physical systems (Mahajan et al., 2020). This is due to the lack of metrics and that the few good ones are often used for applications only marginally related to their original purpose (Novikova et al., 2017). We are not even talking about non-functional requirements here, but just basic descriptions of how well an algorithm performs. Consequently, it is deeply questionable whether the tests conducted to measure the performance of AI will be representative of its performance in real life.

6 Implications for Arms Control

This leads to consideration of implications for arms control. The problems with V&V clearly show that the existing military safety culture will not be sufficient to counter the problems endemic to AI such as safety, security, predictability, control, transparency, trust, brittleness, reliability, etc. These difficulties have been well-documented, so they will only be touched upon briefly. Systems could behave in unexpected ways on the battlefield, especially if they display emergent behavior or employ offline learning. There are also considerable safety risks, because modern weapons are of complex, opaque, interactive, tightly coupled systems, making it hard to identify and put an end to incidents in time. The lack of V&V could decrease operator trust, which is a known risk factor for accidents (Gao et al., 2013). Systems could interact with each other, leading to unforeseen consequences. This is especially risky when it involves systems from opposing parties, as this could lead to unintended conflict escalation. Moreover, the lack of V&V can hide cyber-vulnerabilities and we cannot be certain that the system could not, for example, fall victim to Adversarial AI. Testing would need to take into account the potential behavior of an intelligent adversary in order to ensure system robustness, which is very difficult to accomplish (Tate, 2019a, b). Thus, military systems with AI that cannot be verified and validated pose serious risks for combatants and civilians alike.

This gap in V&V has serious implications for Article 36 reviews. To do legal reviews properly, countries need full access to the technical data of their systems. They must know how the algorithm functions to assess its legality. They will also want to conduct V&V to ensure the system meets the requirements, calibrate trust levels and check for incompatibilities with national doctrines. But AI is often black boxed, making it difficult to conduct legal reviews.

This is even more difficult for arms importers. Arms companies have become very hesitant to share technical data with buyers, even if the countries in question are close allies, under the guise of Intellectual Property Rights (IPR) protection (Langella, 2013). This reluctance is especially strong for weapon systems with AI. French policy makers have also informally expressed their intention to black box all AI arms exports. They are afraid that the algorithms could be reverse engineered and provide information about the highly sensitive data the algorithm was trained on. Importers might assume that exporters will have done legal reviews

for them. However, national doctrine on AI varies. Moreover, only around 20 countries actually conduct Article 36 reviews. The scope of applicable law also varies from country to country, as does the legal interpretation of those laws (Boulanin & Verbruggen, 2017). In addition, Article 36 requires only states to conduct these reviews for their new weapons, means, and methods of warfare. While systematic empirical data is non-existent, off-the-record conversations suggest that military technologies developed by non-states—such as defense companies or the EU—or derivations of weapons meant for export only, do not always undergo Article 36 reviews. Thus, countries cannot assume that weapons are legal and validated and properly verified, and so must conduct TEV&V themselves.

7 The Search for Solutions

Fortunately, defense organizations are aware of these problems and actively searching for solutions. There are many technical solutions currently under investigation, including automated testing, digital twinning, digital test beds, cyber ranges and M&S (Flournoy et al., 2020). But because the literature overwhelmingly focuses on this issue from the point of view of the developer, potential solutions here are discussed from an arms control point of view.

When procuring an externally developed weapon system, provisions regulating IPR, life cycle maintenance, technical data access, maintenance, data ownership, referent data, etc., should be included in all contractual arrangements to avoid situations where countries are unable to inspect the internal workings of the algorithm. In cases where developers refuse to provide information about the internal workings of the system, they should provide a package of relevant information on the specifications, R&D process, data used and all TEV&V results. The extent and quality of available information should be explicitly incorporated into the Article 36 review, which should involve external civilian lawyers.

External and independent verification, validation and certification of weapon systems should become mandatory if it is not already so. The exact regulatory scheme of weapon certification differs from country to country, but it is generally conducted by the industry that developed the weapon, the higher-ranking military officers who favored its adoption, and the MoD. Government bodies such as the FAA, that are supposed to conduct independent certification, are frequently subjected to “regulatory capture,” as they are starved for resources and rely on industry to self regulate. The tragedies with the Boeing 737 MAX were the result (Travis, 2019). Therefore, civilian governmental agencies should receive adequate funding sufficient for them to maintain independent oversight, without any obligations to either the MoD, the military, or the defense industry, and without the tunnel vision that generally accompanies military procurement.

8 Conclusion

This chapter has highlighted the problems with TEV&V of AI. The key point here is that despite all the concerns, it is very difficult to *know* AI, let alone do so with any level of certainty. Integrating AI into weapon systems increases this uncertainty for the entire system. Moreover, our existing TEV&V processes are not ideally suited to correcting these issues. Because of this, we are already seeing political struggles over the standards for TEV&V of AI. If there are no global or national standards to validate and verify AI, developers should instead increase the transparency of algorithms, the development process and the way they have conducted TEV&V of AI. Unfortunately, we are observing a global shift toward less and not more sharing of technical data in weapon systems. But without being able to know how AI will perform, it is doubtful that Article 36 reviews will be sufficient to ensure that no weapons are employed that cannot comply with international law.

References

- Bhattacharyya, S., Cofer, D., Musliner, D., Mueller, J., & Engstrom, E. (2015). *Certification considerations for adaptive systems*. Presented at the 2015 International Conference on Unmanned Aircraft Systems. IEEE. <https://doi.org/10.1109/ICUAS.2015.7152300>
- Bolton, M., Bass, E., & Siminiceanu, R. (2013). Using formal verification to evaluate human-automation interaction: A review. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(3), 488–503. <https://doi.org/10.1109/TSMCA.2012.2210406>
- Boulanin, V., & Verbruggen, M. (2017). *Article 36 reviews: Dealing with the challenges posed by emerging technologies*. SIPRI. Retrieved February 25, 2022, from <https://www.sipri.org/publications/2017/other-publications/article-36-reviews-dealing-challenges-posed-emerging-technologies>
- Braiek, H., & Khomh, F. (2020). On testing machine learning programs. *Journal of Systems and Software*, 164, 110542. <https://doi.org/10.1016/j.jss.2020.110542>
- Brusoni, S., & Prencipe, A. (2011). Patterns of modularization: The dynamics of product architecture in complex systems. *European Management Review*, 8(2), 67–80. <https://doi.org/10.1111/j.1740-4762.2011.01010.x>
- Clark, M. (2015). *Test and evaluation, verification and validation of autonomous systems*. Presented at the Safe and Secure Systems & Software Symposium. AFRL.
- Cook, S., & Haverkamp, G. (2020). Challenges and opportunities for software development and verification on military aircraft systems. In *Scitech 2020 forum*. AIAA. <https://doi.org/10.2514/6.2020-0238>
- Deonandan, I., Valerdi, R., Lane, J., & Macias, F. (2010). *Cost and risk considerations for test and evaluation of unmanned and autonomous systems of systems*. Presented at the 2010 5th International Conference on System of Systems Engineering. IEEE. <https://doi.org/10.1109/SYSOSE.2010.5544062>.
- Dijkstra, E. W. (1972). The humble programmer. *Communications of the ACM*, 15(10), 859–866. <https://doi.org/10.1145/355604.361591>
- Flournoy, M., Chefitz, G., & Haines, A. (2020). *Building trust through testing*. CSET. Retrieved February 25, 2022, from <https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf>

- Gao, F., Clare, A., Macbeth, J., & Cummings, M. (2013). *Modeling the impact of operator trust on performance in multiple robot control*. Presented at trust and autonomous systems. Stanford: AAAI. Retrieved February 25, 2022, from <https://core.ac.uk/display/78055633>
- Gutmann, P. (2004). Verification techniques. In *Cryptographic security architecture: Design and verification* (pp. 84–125). Springer. <https://doi.org/10.1007/b97264>
- Handelman, G., Kok, H.-K., Chandra, R., Razavi, A., Huang, S., Brooks, M., et al. (2019). Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. *American Journal of Roentgenology*, 212(1), 38–43. <https://doi.org/10.2214/AJR.18.20224>
- Haugh, B., Sparrow, D., & Tate, D. (2018). *The status of test, evaluation, verification, and validation of autonomous systems*. IDA. <https://www.jstor.org/stable/resrep22759.1>
- Hylving, L., & Schultze, U. (2020). Accomplishing the layered modular architecture in digital innovation: The case of the car's driver information module. *The Journal of Strategic Information Systems*, 29(3), 101621. <https://doi.org/10.1016/j.jsis.2020.101621>
- Johnson, C. W. (2006). What are emergent properties and how do they affect the engineering of complex systems? *Reliability Engineering & System Safety*, 91(12), 1475–1481. <https://doi.org/10.1016/j.res.2006.01.008>
- Keane, J., & Joiner, K. (2020). Experimental test and evaluation of autonomous underwater vehicles. *Australian Journal of Multi-Disciplinary Engineering*, 16(1), 67–79. <https://doi.org/10.1080/14488388.2020.1788228>
- Langella, F. (2013). *The Italian experience in military type certification within direct national and FMS case acquisition: The M346 and the Predator B*. Presented at the EDA Military Airworthiness Conference, Aix-En-Provence.
- Luckcuck, M., Farrell, M., Dennis, L., Dixon, C., & Fisher, M. (2019). A summary of formal specification and verification of autonomous robotic systems. In W. Ahrendt & S. Tapia Tarifa (Eds.), *Integrated formal methods* (Vol. 11918, pp. 538–541). Springer International Publishing. https://doi.org/10.1007/978-3-030-34968-4_33
- Lyons, J., Clark, M., Wagner, A., & Schuelke, M. (2017). Certifiable trust in autonomous systems: Making the intractable tangible. *AI Magazine*, 38(3), 37–49. <https://doi.org/10.1609/aimag.v38i3.2717>
- Mahajan, V., Venugopal, V., Murugavel, M., & Mahajan, H. (2020). The algorithmic audit: Working with vendors to validate radiology-AI algorithms—How we do it. *Academic Radiology*, 27(1), 132–135. <https://doi.org/10.1016/j.acra.2019.09.009>
- Novikova, J., Dušek, O., Cercas Curry, A., & Rieser, V. (2017). *Why we need new evaluation metrics for NLG*. Presented at the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen: ACL. <https://doi.org/10.18653/v1/D17-1238>
- Pereira, A., & Thomas, C. (2020). Challenges of machine learning applied to safety-critical cyber-physical systems. *Machine Learning & Knowledge Extraction*, 2(4), 579–602. <https://doi.org/10.3390/make2040031>
- Reim, G. (2019, June 6). Optionally piloted Sikorsky UH-60A makes first manned flight. *Flight Global*. Retrieved February 25, 2022, from <https://www.flightglobal.com/helicopters/optionally-piloted-sikorsky-uh-60a-makes-first-manned-flight/133007.article>
- Schaffer, K., & Voas, J. (2016). What happened to formal methods for security? *Computer*, 49(8), 70–79. <https://doi.org/10.1109/MC.2016.228>
- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., et al. (2015). *Hidden technical debt in machine learning systems*. Presented at NIPS 2015, Montreal. <https://doi.org/10.5555/2969442.2969519>
- Tallant, G., Buffington, J., Storm, W., Stanfill, P., & Krogh, B. (2006). *Validation & Verification for emerging avionics systems*. Presented at the National Workshop on aviation software systems: Design for certifiably dependable systems. NITRD. Retrieved February 25, 2022, from https://ptolemy.berkeley.edu/projects/chess/hcssas/papers/Storm-HCSS_avionics_position_paper.pdf
- Tate, D. (2019a). *Attack surfaces of autonomy*. Presented at the 6th cybersecurity workshop challenges facing test and evaluation. ITEA. Retrieved February 25, 2022, from <https://www.itea.org/wp-content/uploads/2019/03/Tate-David.pdf>

- Tate, D. (2019b). *What counts as progress in the T&E of autonomy?* IDA.
- Travis, G. (2019, April 18). How the Boeing 737 Max disaster looks to a software developer. *IEEE Spectrum: Technology, Engineering, and Science News*. Retrieved February 25, 2022, from <https://spectrum.ieee.org/how-the-boeing-737-max-disaster-looks-to-a-software-developer>
- Varshney, K. R., & Alemzadeh, H. (2017). On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big Data*, 5(3), 246–255. <https://doi.org/10.1089/big.2016.0051>
- Wayne, H. (2019, January 21). Why don't people use formal methods? *Hillel Wayne*. Retrieved February 25, 2022, from <https://hillelwayne.com/post/why-dont-people-use-formal-methods/>
- Wojton, H., Porter, D., & Dennis, J. (2020). *Test & evaluation of AI-enabled and autonomous systems: A literature review*. IDA. Retrieved February 25, 2022, from <https://testscience.org/wp-content/uploads/sites/16/formidable/20/Autonomy-Lit-Review.pdf>
- Young, S. (2016). *Autonomy test & evaluation verification & validation challenge area*. Presented at the 31st National Test & Evaluation Conference, McLean. Retrieved February 25, 2022, from <https://ndiastorage.blob.core.usgovcloudapi.net/ndia/2016/Test/Young.pdf>

Further Reading

- Haugh, B., Sparrow, D., & Tate, D. (2018). *The status of test, evaluation, verification, and validation of autonomous systems*. IDA. <https://www.jstor.org/stable/resrep22759.1>
- Helle, P., Schamai, W., & Strobel, C. (2016). Testing of autonomous systems - Challenges and current state-of-the-art. *INCOSE International Symposium*, 26(1), 571–584.
- Wojton, H., Porter, D., & Dennis, J. (2020). *Test & evaluation of AI-enabled and autonomous systems: A literature review*. IDA. Retrieved February 25, 2022, from <https://testscience.org/wp-content/uploads/sites/16/formidable/20/Autonomy-Lit-Review.pdf>

Applying Export Controls to AI: Current Coverage and Potential Future Controls



Kolja Brockmann

● آقای هوش مصنوعی ●

رسانه هوش مصنوعی دانشگاه تهران

@MrArtificialintelligence

Abstract Advances in artificial intelligence (AI) are increasingly conflated with military power and competition has ensued over technological leadership. States are looking to export controls as an instrument to prevent misuse of AI and govern the trade in relevant technologies. However, there is a lack of clarity about the extent to which export control instruments already cover dual-use goods and technologies used in AI and its military applications. A growing debate is emerging in national and in multilateral forums about both expanding export controls and about their limitations in the context of AI. This chapter reviews the role, coverage and limitations of export controls, explores recent initiatives to expand export controls and identifies opportunities for export controls to contribute to a comprehensive governance approach to AI.

1 Introduction

Artificial intelligence (AI) is an umbrella term that was originally coined in the 1950s and is now commonly used to describe a range of computational techniques that allow machines to perform tasks and carry out deliberation processes that are usually associated with human intelligence, including natural language processing, computer vision and learning (Boulanin, 2019, pp. 13ff).¹ Advances in AI are bound to play a significant role in the future development and capability of conventional weapons, nuclear weapons, advanced delivery systems, cyber tools and surveillance technologies (Boulanin et al., 2020a, b; Brockmann et al., 2019; Viski et al., 2020). Advances in AI are increasingly conflated with military power and superiority and a fierce competition has ensued, particularly between the United States (US) and

¹This chapter only considers so-called *weak* or *specialized* AI, in contrast to *strong* AI or what is commonly referred to as artificial general intelligence (AGI).

K. Brockmann (✉)

Stockholm International Peace Research Institute (SIPRI), Stockholm, Sweden

e-mail: kolja.brockmann@sipri.org

China, but increasingly also involving European Union (EU) member states, Russia and other states. Concerns about the potential humanitarian and strategic impact of advances in AI and their adoption by militaries are driving a search for hard and soft law instruments to prevent misuse of AI and related applications, encourage restraint, promote transparency and build confidence. The current and future role of export controls in strengthening the oversight and regulation of the trade in and use of AI is gaining more and more attention in this wider debate. However, much of the discussion is characterized by uncertainty about the current and potential future capabilities of AI technology and the role that export controls can play in reducing its risks.

There is a lack of clarity about the extent to which export control instruments already cover dual-use goods and technologies used in AI and its military applications. There is also a growing debate both nationally and in multilateral forums about the possible need for additional export controls and about what form they should take (Barkin, 2020; Flynn, 2020; Viski et al., 2020). Much of the debate is focused on the challenges and adverse consequences of potential new controls, including the difficulty of identifying specific AI technologies that could usefully be covered by export controls and the potential negative impact of such controls on AI research and development.

This chapter seeks to improve understanding of the role of export controls in governing the trade in and use of AI technology and provides a preliminary assessment of their current and future role. In doing so, it seeks to amplify on ongoing discussions about the potential need for and risks of adopting new export control instruments in this area. Section 2 introduces export controls and their role as a non-proliferation and technology governance tool. It provides an overview of the existing coverage of AI and its military and dual-use applications by different export control instruments. Section 3 discusses recent national and multilateral initiatives to review potential new export controls on AI. Section 4 analyzes the main challenges and limitations of applying export controls to AI. It also considers specific opportunities that export controls and related compliance measures provide for the governance of AI. Finally, Sect. 5 briefly summarizes findings and policy recommendations.

2 Current Export Controls on AI and Related Hardware, Software and Technology

Export controls are sets of regulations established by states to gain oversight over and limit trade in certain military and dual-use goods and technologies.² They impose licensing requirements—and in some cases prohibitions—on transfers of

²Contrary to the common understanding of *technology* as “machinery and equipment developed from the application of scientific knowledge” (Concise Oxford English Dictionary, 2022), in export

controlled goods and technologies. Export controls establish rules for when export licensing applications may be granted or denied by a state, on what basis and according to which standards and criteria these decisions should be taken. The implementation of export controls is reliant on control lists that define those goods and technologies to which licensing requirements apply, and guidelines based on which end uses and end users should be restricted. The control lists and guidelines adopted by most states implementing export controls have been created and are maintained by multilateral export control regimes. The four main regimes are the Australia Group (covering chemical and biological weapons), the Missile Technology Control Regime, the Nuclear Suppliers Group (NSG) and the Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies (WA). They are informal groups of supplier states that seek to harmonize their export controls through politically binding guidelines to prevent destabilizing accumulations of conventional weapons and the proliferation of chemical, biological and nuclear (CBN) weapons and their delivery systems. Despite the regimes' exclusive membership and thus limited participation in their deliberation and decision making, their control lists and guidelines have been adopted or adapted by many states outside their membership (Brockmann, 2019). Notably, some major exporters remain outside some of the regimes. For example, China only participates in the NSG, and Israel does not participate in any of the regimes, although both are major exporters of unmanned aerial vehicles and many other military and dual-use goods and technologies.

The WA is the most relevant export control regime in the context of AI because of its coverage of conventional arms and dual-use goods and technologies. The guidelines and control lists of the WA (Wassenaar Arrangement, 2019) are the basis for most national export control regulations in this area, including in the EU and the US. The EU dual-use regulation—which is directly applicable in EU member states—combines the control lists of the four multilateral export control regimes in one list (Council of the European Union, 2021). The EU military list is also based on the WA munitions list (Council of the European Union, 2019). The US export control system implements the WA guidelines and control list, but some of its regulations go beyond them, particularly through the extraterritorial application of US re-export controls and the application of controls on so-called *deemed* exports and re-exports.³

Export control regulations—but also the related trade facilitation mechanisms—impose certain levels of due-diligence requirements on exporting companies or research institutes. Thus, export control compliance is not only a requirement for exporting companies but also for researchers, scientists and academics—among

control regulations *technology* is usually defined as “specific information necessary for the ‘development,’ ‘production’ or ‘use’” of an export-controlled item (Wassenaar Arrangement, 2019, p. 234).

³A deemed export is the release of controlled technology or software in the US to a national of another country. A deemed re-export is the release of controlled technology or software “in one foreign country to a national of another foreign country” (US Department of Commerce, 2020).

others—who share and transfer knowledge, data and technology that could be used to develop, produce or use controlled items to foreign nationals or to persons or entities abroad. Procedures for ensuring compliance, due diligence, ethical research practices and awareness among researchers, scientists, engineers and commercial companies are therefore often components of, or complement, effective export control governance frameworks. Other complementary measures include foreign direct investment controls, visa screening, targeted sanctions, embargoes, diplomacy, and other policy tools that seek to influence individuals' and states' behavior.

2.1 Key Types of Export Controls on AI

There is currently a range of both list-based and end-use/r-based export controls that already apply to certain AI hardware, software, technology and their specific applications. Some of these controls are incidental and overlapping controls that result from controls that were introduced in another context in the past, but are still relevant for contemporary AI development, production and use. In addition, the way that export controls on software and technology are commonly designed means that they—in many cases—already cover AI software and technology that is *specifically designed* or *modified* for use in military systems or for end uses in CBN weapons or their delivery systems. However, much of the research and development in the area of AI will continue to be multi-purpose, widely published in open-source literature, frequently shared among AI researchers worldwide, and not specific to these end-uses (OpenAI, 2019; Stanley-Lockman, 2019). As a result, it will be difficult to govern advances and access to these more general AI technologies using export controls. The main types of export controls relevant to the control of AI are hardware, software, technology and catch-all controls.

2.1.1 Hardware Controls

The WA's dual-use control list—largely adopted by the EU and the US—includes several specific list items covering AI-related hardware, including neural network integrated circuits, neural computers, a range of different sensors, several types of computer chips and some production equipment for semi-conductors.⁴ However, none of these represent what is commonly referred to as *technological choke points* without which the development or acquisition of specific AI capabilities by a third country would be seriously impeded or prevented. Identifying, defining and controlling specific items that are indispensable for the technology but not widely available are key if access to specific technologies is to be denied by applying export controls.

⁴For example, see items controlled under 3.A.1.a.9., 4.A.4.b. and Category 6 on the WA control list.

Some relevant items are listed—for example “*neural computers*”—but they fail to impose meaningful controls (Thomsen, 2018, pp. 16 f.). The formulation of some of these list items no longer reflects the terminology and state of the art of the technology or their technical parameters do not correspond to current developments. Other AI-specific hardware, such as chips that are specifically designed to enable higher computing power in the training of AI models, or specific manufacturing equipment for their production, is currently not covered by any specific list item on the multilateral or any national control lists (Flynn, 2020, p. 8).

2.1.2 Software Controls

Export controls apply to the transfer and making-available of a wide range of software. Dual-use control lists do not usually list software in a separate category but, instead, several control list items include controls on software which is “*pecially designed or modified*” for the “*development, production or use*” of the controlled item.⁵ While most software controls are thus linked to a specific controlled item, there is also some dual-use software that is controlled in and of itself. Software that is generally available to the public through retail without restrictions or is “in the public domain” (Wassenaar Arrangement, 2019, p. 3) is exempt from controls. Much of the basic AI software that is already widely shared and published *open access* is thus exempt. The WA’s munitions lists (ML), which cover military goods and technologies, include a separate category—ML21—for controlled software. ML21 covers software “designed or modified” either for the “development, production and maintenance” of equipment and software specified on the munitions list, or for the “development or production” of any material covered by the munitions list (Wassenaar Arrangement, 2019, p. 212). Trained models and specific AI systems—particularly those that can be used for the decision and action steps in specific AI applications—are usually highly application-specific and their end use, and thus whether they are subject to licensing requirements, is much easier to discern. For example, AI software that has been developed or modified for use in automated close-in weapons systems or air defense systems would require an export authorization. The US is currently the only state that has introduced unilateral export controls on specific AI software. In January 2020, the US Department of Commerce added a control list item covering “*Geospatial imagery “software” “pecially designed” for training a Deep Convolutional Neural Network to automate the analysis of geospatial imagery and point clouds*” to the Commerce Control List ECCN 0Y521 series (US Department of Commerce, Bureau of Industry and Security, 2020a)—a special category for temporary controls on items not previously

⁵On the dual-use control list of the Wassenaar Arrangement, each control list category has a “Subcategory D” specifying those listed items for which software as defined is controlled.

listed, particularly emerging technologies (US Department of Commerce, Bureau of Industry and Security, 2020b).⁶

2.1.3 Technology Controls

As defined by the multilateral export control regimes and adopted in EU and US export control regulations, technology describes “specific information necessary for the ‘development,’ ‘production’ or ‘use’ of a product” (Wassenaar Arrangement, 2019, p. 234).⁷ It can take the form of *technical data* and *technical assistance*. The WA’s dual-use control list specifies that *technical data* can be “blueprints, plans, diagrams, models, formulae, tables, engineering designs and specifications, manuals and instructions” (Wassenaar Arrangement, 2019, p. 234). It defines *technical assistance* as activities such as “instruction, skills training, working knowledge, consulting services,” including when these “involve transfer of ‘technical data’” (Wassenaar Arrangement, 2019, p. 234). In a manner similar to controls on military software, controls on technology have their own control list category on the WA’s munitions list—ML22. These controls largely follow the General Technology Note of the dual-use list, but also extend to specific technology for production installations (Wassenaar Arrangement, 2019, see ML 22.b.1). Technology controls extend to transfers or making available of datasets used to train a specific AI system that is specially designed for a listed military item or if the information contained in the dataset constitutes technical data as defined above. However, making this distinction is inherently difficult, particularly for AI systems and datasets that are for general purposes and less application-specific. Moreover, export controls on technology also include exemptions for transfers of items that constitute *basic scientific research* or *information in the public domain* (Wassenaar Arrangement, 2019, p. 3). Such exemptions would probably apply in the case of much basic AI technology.

2.1.4 Catch-All Controls

Catch-all controls are a common export control instrument that enables states to apply controls on transfers of non-listed dual-use goods and technologies if available information suggests that they are destined for a restricted end use or end user. The application of catch-all controls has been discussed extensively in the context of emerging technologies, including AI (Viski et al., 2020, pp. 40, 52). Catch-all controls can be a useful tool to enable states to balance security-driven needs for

⁶For a more detailed discussion of the use of the ECCN 0Y521 series to control emerging technologies see Brockmann (2018, pp. 20–22).

⁷The term *development* relates to “all stages prior to serial production”; “production” refers to “all production stages”; and “use” refers to “operation,” “installation” and “maintenance” of an item (Wassenaar Arrangement, 2019, pp. 219, 228, 235).

more control with economically-driven free trade and competitiveness imperatives. They enable a state to impose controls without introducing broad, list-based controls that would increase the burden on exporters and licensing authorities to apply for and process more licensing applications (Brockmann & Kelley, 2018, pp. 25–26; Bromley & Bauer, 2016, p. 7). However, national authorities are often dependent on access to intelligence information and the due-diligence procedures of companies to identify cases in which they can apply catch-all controls. Extensively relying on catch-all controls can create uncertainty about whether controls apply, especially when compared with controls based on list items with specific technical parameters (Bromley & Bauer, 2016, p. 7). The participating states in the WA, including the US and the EU member states, agreed to also establish catch-all provisions on transfers of non-listed dual-use items with military end uses. However, these controls are limited to transfers to “destinations subject to a binding United Nations Security Council arms embargo” (Wassenaar Arrangement, 2003, p. 1) and relevant regional arms embargoes that are binding or adhered to by the exporting state (Wassenaar Arrangement, 2003; Bromley & Maletta, 2018, pp. 17–18).⁸ The participating states reserved the right to adopt *national measures* to restrict exports for other reasons of public policy, enabling some variation in specific national catch-all controls on dual-use items with military end uses.

3 National and Multilateral Approaches to Expanding Export Controls on AI

Several national and multilateral review processes have recently been initiated to assess and develop recommendations for expanding export controls and related strategic trade management tools on emerging technologies, including AI. The three most significant initiatives and processes have been initiated by the US, the WA and the EU and illustrate the considerations taken into account in this area.

3.1 US Emerging Technologies Review Process

In 2018, the United States initiated a public consultation process to identify “specific emerging technologies that are essential to the national security of the United States” (US Department of Commerce, Bureau of Industry and Security, 2018, p. 2). Major US and multinational companies cautioned that the technology areas identified in the

⁸The Wassenaar Arrangement agreed in 2003 that catch-all controls should apply to transfers to “destinations subject to a binding United Nations Security Council arms embargo, any relevant regional arms embargo either binding on a Participating State or to which a Participating State has voluntarily consented to adhere” (Wassenaar Arrangement, 2003, p. 1).

AI and machine learning category are long-established and pointed out that many “have been subject to dual-use export controls for over 20 years” (IBM, 2019, p. 7). They further cautioned against a potential negative impact of restrictive controls on the economy and security, loss of technology leadership, access to talent, and diminished access to the best technology by both industry and the military (TradeSecure, 2019).

In parallel to the public consultation, the US government has been engaged in an interagency review process to explore potential controls on AI-relevant hardware, software and technology. In terms of hardware that could potentially be controlled, many options that could be defined by technical parameters have been considered, but no agreement has been reached on introducing controls on specific ones yet.⁹ For example, controls on chips designed to accelerate the computing required in AI systems are being considered and have been suggested by a number of analysts. These chips could be clearly defined using technical parameters and restricting access to them could have a significant non-proliferation effect concerning advanced AI capabilities by making their acquisition and deployment more difficult (Kania, 2017; Stanley-Lockman, 2019).

3.2 EU Emerging Technologies Review and Dual-Use Regulation Recast

In 2019 and 2020, the European Commission held a series of *Technical Workshops on Emerging Technologies*, bringing together the European Commission and all interested EU member states for discussions among government technical experts on a number of emerging technologies. Several such workshops took place between November 2019 and December 2020, including some in a virtual format due to the COVID-19 pandemic (European Commission, 2021). One of the workshops discussed the topic of AI, but it did not result in any immediate initiatives to propose new AI-related controls in the regimes.¹⁰

The EU has been engaged in a process of reviewing and recasting the EU Dual-Use Regulation since 2011 (Bromley & Maletta, 2019). One of the changes considered was the introduction of an *EU-autonomous list*—meaning a control list independent of the multilateral export control regimes—through which the EU could introduce controls binding on all member states, for example on emerging technologies. However, the recast regulation that entered into force in September 2021 instead created transmissible controls through which one EU member state can apply

⁹The author participated in a dialogue meeting on AI and strategic trade controls in March 2020 with participants from the relevant departments of the US government and conducted background interviews with senior officials from the US Department of Commerce.

¹⁰Interviews conducted by the author with national licencing officials involved in the review process.

catch-all controls to exports of non-listed items for which a national control list entry has been created by another member state (Bromley & Brockmann, 2021). This compromise reflects the EU's continued position that controls should be agreed in the regimes but acknowledges that for some emerging technologies timely unilateral controls by member states may be required.

3.3 Wassenaar Arrangement Discussions on Export Control on AI

In 2017, the WA's Head of Secretariat said that "Looking ahead, the WA Lists review process can be expected to continue to address new technologies of security concern, including [...] artificial intelligence and the integration of advanced sensors and navigation equipment to increase autonomy of weapons systems" (Griffiths, 2017, p. 3). By the end of 2021, no new AI-related control list items or amendments to existing controls had been agreed upon and published by the participating states.

4 Challenges and Opportunities of Applying Export Controls to AI

The wider policy debate and the ongoing review processes on the potential expansion of export controls on AI have revealed a range of challenges to and adverse consequences of applying export controls to AI. They have however also identified some areas where the application of export controls promises benefits and presents opportunities for strengthening governance approaches to protect international peace and security while fostering responsible development and innovation in AI research and industry.

4.1 Challenges and Adverse Consequences

4.1.1 Implementation Challenges

Several inherent, mainly technical limitations affect the effective application and sustainability of export controls on AI. There is a lack of technological choke points that would allow key AI technologies to be identified and defined, without which the development of sensitive AI-enabled systems would be impossible or at least considerably impeded (Thomsen, 2018, pp. 15 f.). In addition, the speed of development of AI technology and technical parameters and performance levels of AI systems makes it difficult to identify and adjust control list items quickly enough, particularly through consensus decisions in the multilateral regimes

(Brockmann, 2018). The established practices of the AI research community, including the sharing of algorithms, data, models and research findings more broadly—as is the case for many dynamic fields of dual-use research, including the life sciences and cybersecurity (Shaw, 2016; Bromley, 2017; Hinck, 2018)—are difficult to balance with the increasing pressure to impose regulations and governance tools, including export controls. One serious difficulty here is that the specific systems that are of concern are often not clearly identified, whether because their development is proprietary, they are yet to be developed and refined, or because issues over what by definition constitutes specific AI systems of concern have not been sufficiently addressed. To date, the different review processes have not been able to identify technical parameters for specific AI applications that would otherwise not be covered, but should be, and sufficiently distinguish them from other civilian applications.

The implementation of existing controls on intangible transfers of technology (ITT) and non-list-based controls, such as catch-all controls, also poses a number of challenges (Stewart, 2016). In the case of AI, there is strong reliance on ITT, either through electronic transfers of software and technology or through the transfer of intangible and often tacit knowledge in the form of teaching, international cooperation or work with foreign nationals. Export controls on software and technology are criticized by some from the policy analyst community as outdated for reasons including their not being physical goods crossing borders that could be inspected and stopped (Leung et al., 2019). However, the digitization of information and the implementation challenges of controls on ITT and software have been topics discussed in the multilateral regimes for over a decade and states have increasingly taken steps to adjust and improve controls (Brockmann, 2019, p. 14). Detection and enforcement of controls on ITT stand out as areas where notable difficulties persist (Bauer & Bromley, 2019). It is often a difficult process to establish what is subject to control and what is not, because there is a lack of clarity on what is *required* and constitutes *production*, *development*, and *use*. However, the same problems apply to tangible dual-use goods. While the enforcement of ITT controls continues to be challenging, steps have been taken by some states to strengthen enforcement, including the implementation of specialized audit procedures using digital forensics techniques to verify digital record keeping requirements and trace sharing and transfers of controlled data or other information (Bauer & Bromley, 2019). As ITT remains an area of export controls that is particularly reliant on cooperation from exporters, efforts are underway in many states to raise awareness and strengthen compliance and self-regulatory practices in academia, research institutes and research and development departments of companies.

Controls based on end use and end users depend on the information and intelligence available to companies, universities, research institutes and states. Thus, to a significant extent they are a function of the intelligence gathering abilities and access of states, as well as the quality of due-diligence procedures and awareness among exporters. In both areas there are significant disparities among states and in different sectors of industry and research, many of which are either using AI or are involved in relevant supply chains (Bauer et al., 2017). Outreach and awareness-raising programs and harmonization and exchange of good practices in the implementation of

non-list-based controls with regard to the field of AI are therefore advisable. Catch-all controls may offer an interim possibility for states to apply controls, particularly in the case of WMD end uses, but their individual application and the uncertainty that their application creates for research and industry make them less suitable in the longer run.

Another challenge to the effective implementation of export controls on AI is the implementation of due-diligence procedures and carrying out technology classifications by companies, research institutes and universities. This is particularly the case if the technology or product developed is still somewhat removed from the final product and end use. Traditional export control measures should thus be complemented by dedicated awareness raising and engagement programs throughout the supply chains that can, for example, contribute to the development of autonomous weapon systems (AWS), facial recognition systems and other AI applications that could be misused for repression and other human rights or international humanitarian law (IHL) violations. Many large companies have shown they are aware of and are even struggling internally with their employees over the role they are taking in the development of AI technologies and their applications (IBM, 2020; Rasser et al., 2019, pp. 21ff.). Many smaller companies in relevant supply chains may be more susceptible to incentives offered by critical end users or lack the compliance structures to prevent inadvertent transfers of technology.

4.1.2 Potential Adverse Consequences

Companies and researchers, particularly those from Silicon Valley in the US, have extensively warned of potential adverse consequences of expanding export controls on AI (Metz, 2019). One particular concern pertains to a potential reduction of international cooperation in AI development as a result of stricter export controls. If such stricter policies were to extend to more restrictive visa screening and controls on deemed exports, they could also have a significant impact on access to and retention of AI talent (Leung et al., 2019). Companies such as Google have argued that “expansive controls on AI technology” could affect a significant number of their engineers and that even if licenses were to be commonly granted, the perception of a constrained innovation environment would affect their ability to retain talent and consequently slow the rate of new product development (Google, 2019, pp. 8 f.).

While many concerns have centered around economic impact and competitiveness, a more restrictive and nationalized environment for AI development could also reduce states’ willingness to be transparent and share information about their domestic AI development and adoption in military and security applications. This could be particularly relevant in light of the competition dynamics that are already playing out today in the civilian realm and may spill over in the military realm as capabilities and usability of AI systems increase. Finally, this could mean that the development of AI-enabled military technologies might be affected, including less willingness on the part of AI companies and researchers to participate in defense product development to avoid being affected by restrictive export controls.

Particularly those states pursuing ambitious strategies for the development of military AI are sensitive to such effects and are likely to take them into consideration when developing export controls in the area of AI. States in favor of limits on military AI development, on the other hand, may see this as a positive effect that encourages responsible practices and the exercise of restraint on the part of AI researchers and industry.

4.1.3 Conflicting Aims of Export Controls on AI

A complicating issue concerns the different aims pursued by export controls on AI. The most straightforward aim is preventing the proliferation of AI technology and its application in, or linked to, CBN weapons and their delivery systems. However, in the area of conventional weapons systems, where post-Cold War export controls have aimed at strengthening transparency and restraint to prevent destabilizing accumulations of weapons (Wassenaar Arrangement, 2011), using export controls in the context of AI is more complicated. For example, with regard to autonomy in weapons systems there is currently no ban, specific regulation of capabilities, or required degree of human control in place. Weapons systems must nonetheless comply with and be used in compliance with IHL.¹¹ There are existing requirements to address human rights and IHL concerns in national export controls, including through the EU Common Position on Arms Exports and the guidelines of the WA. Operationalizing these requirements (as well as other considerations emerging from the ongoing debate on lethal autonomous weapons systems) for export control risk assessment procedures and licensing decision-making may nevertheless be difficult. This is particularly challenging in an environment that is strongly marked by competition between the US, China, Russia and other states while at the same time technological advances in AI are conflated with economic and military advantage and power (Barkin, 2020). Notably, a considerable portion of the literature focuses on preserving national technology leadership in AI (e.g., Rasser et al., 2019). This reflects some of the difficulty that is created when seeking to use one instrument to achieve both the multilateral aims of preventing threats to international peace and security, human rights and IHL and the national pursuit of economic and technological advantage, including military-technological superiority, as part of broadly defined national security objectives.

There is no agreed definition on what constitutes *misuse* of AI research and development or on which transfers at the different stages in an AI system's development and life cycle pose risks to international peace and security. Even for the integration of autonomy in weapons systems these risks have not been specifically codified to date and states rely on the interpretation of human rights and IHL

¹¹ Article 36 of the 1977 Additional Protocol to the 1949 Geneva Conventions creates an obligation on states to implement reviews of new weapons systems to ensure their use would not be prohibited by international law (Boulanin & Verbruggen, 2017).

provisions in national risk-assessment procedures as part of decision-making about licensing (iPRAW, 2020, p. 2). An international legal reference that could potentially emerge from the conclusion of an international arms control agreement governing autonomy in weapons systems could enable more coherent and harmonized use of export controls in this context. The absence—to date—of such a treaty or specific normative system limits the ability to discriminate appropriately and apply a harmonized standard in export controls with respect to autonomy in weapons systems. Beyond the specific risks related to autonomy, the wider military use of AI, including in battle-management and decision-support systems, poses a host of other risks that would also need to be part of risk assessments in research and development and in export licensing. In practice, there should be a strong presumption of denial of transfers of AI-related goods and technologies with an end use in fully autonomous weapons systems as well as in repression and other human rights violations. Similarly, states should deny license applications where the receiving state is unable to demonstrate that it will maintain appropriate levels of human control in the operation of autonomy in weapons systems.

4.2 *Opportunities and Benefits*

Imposing export licensing requirements, while ensuring as much legal clarity and predictability as possible, can allow states to increase the oversight over and awareness of transfers of AI hardware, software and technology, even if no license-denials are being issued. This increases the oversight of what companies are developing and marketing as well as who they are supplying and where those being supplied are located. Most importantly, it allows states to more readily limit transfers of sensitive AI-related products. While export controls undoubtedly introduce some adverse effects in a highly competitive field such as AI, they nevertheless provide for a system that contributes to ensuring peaceful uses of transfers of sensitive items through licensing and transparency, rather than pursuing a strategy of technology denial (Evans, 2014, pp. 4 f.). States adopting unilateral export controls would face many of the adverse consequences outlined above, particularly those related to the competitiveness of domestic industry and research. The adverse consequences of stricter export controls on the global AI industry would be particularly significant if such controls were applied by the US, because of the extraterritorial reach of US controls and deemed export controls. In contrast, multilateral controls, for example through the WA, are much more likely to maintain a level playing field in terms of economic competitiveness, as they set common standards, including beyond their membership. Thus, where possible, states should seek to establish new controls on AI through multilateral frameworks.

Some analysts have argued that claims of *AI democratization*—asserting that AI technology has become so ubiquitous and easy to access that imposing barriers at this point would be futile—are exaggerated, particularly in the realm of military applications of AI, as many militaries will not be able “to build up the talent,

computing power and data, and organizational capacity required to sufficiently scale up their usage of AI to produce appreciable effects” (Stanley-Lockman, 2019). As noted by many government officials, the existing export control architecture already offers considerable coverage of military end uses and extensive coverage of all CBN weapons and delivery systems-related transfers. In addition, some hardware that is particularly relevant in identifiably military applications and sufficiently distinguished by technical parameters could also be effectively controlled, for example, hardened and specially designed electronics components and assemblies, as well as specific sensors.

While list-based AI hardware controls and mainly end-use dependent software and technology controls will not be enough to control the proliferation of sensitive applications of AI and govern its development, production and use, they still provide a valuable governance tool. This is particularly the case if they are applied in a comprehensive approach in combination with outreach to researchers and industry, raising awareness with developers, establishment of standards for responsible innovation and ethical research, and other upstream compliance and self-governance mechanisms (Boulanin et al., 2020a). The result could at least be a basic level of control over the proliferation of, increased information on, and oversight over the trade in sensitive AI-related goods and technologies, and increased transparency, particularly among like-minded states, and could also create more widely-adopted standards and norms.

5 Conclusion

Export controls can play a significant role in the oversight and regulation of the trade in and use of AI. Despite only a small number of specific control list items covering multi-purpose AI hardware, software and technology, current export controls already cover a significant range of transfers with military and CBN weapons end uses. Review processes to potentially expand export controls on AI and strengthen related governance and compliance tools are ongoing in the US, the WA and the EU. Blanket controls on AI-related goods and technologies would negatively affect scientific and technological development and prevent reaping the likely benefits of further advances in AI and should thus be avoided. National governments—particularly through multilateral export control regimes coordinated among EU member states and allies—should continue their review processes to identify specific AI-related items that could be covered by limited and straightforward controls and seek to clarify and deconflict the aims pursued by export controls on AI as far as possible. At the same time, states should continue to strengthen complementary standards for compliance programs, awareness raising, responsible innovation, ethical research and self-governance. Export controls and related compliance measures are only one component in the larger governance framework that is required to limit the potential negative impact of AI on international peace and security, human rights and IHL. Coordination and exchange with a wide range of stakeholders,

particularly on other governance approaches, including traditional arms control and responsible research and innovation, will therefore be key in achieving progress going forward.

References

- Barkin, N. (2020). *Export controls and the US-China Tech War: Policy challenges for Europe*. MERICS Mercator Institute for China Studies. <https://www.merics.org/en/china-monitor/export-controls-and-the-us-china-tech-war>.
- Bauer, S., & Bromley, M. (2019). *Detecting, investigating and prosecuting export control violations: European perspectives on key challenges and good practices*. SIPRI. https://www.sipri.org/sites/default/files/2019-12/1912_sipri_report_prosecuting_export_control_violations_0.pdf
- Bauer, S., Brockmann, K., Bromley, M., & Maletta, G. (2017). *Challenges and good practices in the implementation of the EU's arms and dual-use export controls: A cross-sector analysis*. SIPRI. https://www.sipri.org/sites/default/files/2017-07/1707_sipri_eu_duat_good_practices.pdf
- Boulanin, V., Brockmann, K., & Richards, L. (2020a). *Responsible artificial intelligence research and innovation for international peace and security*. SIPRI. https://www.sipri.org/sites/default/files/2020-11/sipri_report_responsible_artificial_intelligence_research_and_innovation_for_international_peace_and_security_2011.pdf
- Boulanin, V., Saalman, L., Topychkanov, P., Su, F., & Peldán Carlsson, M. (2020b). *Artificial intelligence, strategic stability and nuclear risk*. SIPRI. https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf
- Boulanin, V. (Ed.). (2019). *The impact of artificial intelligence on strategic stability and nuclear risk, Vol 1: Euro-Atlantic perspectives*. SIPRI. <https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf>
- Boulanin, V., & Verbruggen, M. (2017). *Mapping the development of autonomy in weapon systems*. SIPRI. https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf
- Brockmann, K. (2019). *Challenges to multilateral export controls: The case for inter-regime dialogue and coordination*. SIPRI. https://www.sipri.org/sites/default/files/2019-12/1912_regime_dialogue_brockmann.pdf
- Brockmann, K. (2018). Drafting, implementing, and complying with export controls: The challenge presented by emerging technologies. *Strategic Trade Review*, 4(6), 5–27.
- Brockmann, K., Bauer, S., & Boulanin, V. (2019). *Bio plus X: Arms control and the convergence of biology and emerging technologies*. SIPRI. https://www.sipri.org/sites/default/files/2019-03/sipri2019_bioplusx_0.pdf
- Brockmann, K., & Kelley, R. (2018). *The challenge of emerging technologies to non-proliferation efforts: Controlling additive manufacturing and intangible transfers of technology*. SIPRI. https://www.sipri.org/sites/default/files/2018-04/sipri1804_3d_printing_brockmann.pdf
- Bromley, M. (2017). *Export controls, human security and cyber-surveillance technology: Examining the proposed changes to the eu dual-use regulation*. SIPRI. https://www.sipri.org/sites/default/files/2018-01/sipri1712_bromley.pdf
- Bromley, M., & Brockmann, K. (2021). Implementing the 2021 recast of the EU dual-use regulation: Challenges and opportunities. EU Non-Proliferation and Disarmament Consortium Paper 77. SIPRI. https://www.sipri.org/sites/default/files/2021-09/eunpdc_no_77.pdf
- Bromley, M., & Maletta, G. (2019). Developments in the European Union's dual-use and arms trade controls. In *SIPRI yearbook 2019: Armaments, disarmament and international security* (pp. 532–537). Oxford University Press.

- Bromley, M., & Maletta, G. (2018). *The challenge of software and technology transfers to non-proliferation efforts: Implementing and complying with export controls*. SIPRI. https://www.sipri.org/sites/default/files/2018-04/sipri1804_itt_software_bromley_et_al.pdf
- Bromley, M., & Bauer, S. (2016). The dual-use export control policy review: Balancing security, trade and academic freedom in a changing world. *EU Non-Proliferation Consortium Paper 48*. SIPRI. <https://www.sipri.org/publications/2016/eu-non-proliferation-papers/dual-use-export-control-policy-review-balancing-security-trade-and-academic-freedom-changing-world>
- Concise Oxford English Dictionary*. (2022). 12th ed. Oxford University Press.
- Council of the European Union. (2019). Council common position 2008/944/CFSP of 8 Dec. 2008 defining common rules governing control of exports of military technology and equipment. *Official Journal of the European Union*, L335, 8 Dec. 2008. Amended by Council Decision (CFSP) 2019/1560 of 16 Sept. 2019. *Official Journal of the European Union*, L239, 17 Sept. 2019.
- Council of the European Union. (2021). Regulation (EU) 2021/821 of the European Parliament and of the Council of 20 May 2021 setting up a community regime for the control of exports, transfer, brokering and transit of dual-use items (recast). *Official Journal of the European Union*, L206, 11 June 2021.
- European Commission. (2021). *Emerging technologies: Developments in the context of dual-use export controls*. Fact sheet. https://trade.ec.europa.eu/doclib/docs/2021/september/tradoc_159791.pdf
- Evans, S. A. W. (2014). *Revising export control lists*. Report. Flemish Peace Institute. https://vlaamsvredesinstituut.eu/wp-content/uploads/2015/03/revising_export_control_lists_web.pdf
- Flynn, C. (2020). Recommendations on export controls for artificial intelligence. *CSET Issue Brief*. <https://cset.georgetown.edu/wp-content/uploads/Recommendations-on-Export-Controls-for-Artificial-Intelligence.pdf>
- Google. (2019). Comments submitted in response to ‘Advance notice of proposed rulemaking: Review of controls for certain emerging technologies’. *Federal Register*, 83(223) (19 Nov. 2018). RIN 0694-AH61.
- Griffiths, P. (2017). The proliferation threat landscape in 2017 – Mounting dangers? WMD/military proliferation trends and emerging technologies of concern. *2017 Export Control Forum*. http://trade.ec.europa.eu/doclib/docs/2017/december/tradoc_156485.pdf
- Hinck, G. (2018, January 5). Wassenaar export controls on surveillance tools: New exemptions for vulnerability research. *Lawfare*. <https://www.lawfareblog.com/wassenaar-export-controls-surveillance-tools-new-exemptions-vulnerability-research>.
- IBM. (2020, September 11). A precision regulation approach to controlling facial recognition technology exports. *IBM THINKPolicy Blog*. <https://www.ibm.com/blogs/policy/facial-recognition-export-controls/>
- IBM. (2019). Comments submitted in response to ‘Advance notice of proposed rulemaking: Review of controls for certain emerging technologies’. *Federal Register*, 83(223) (19 Nov. 2018). RIN 0694-AH61.
- iPRAW (International Panel on the Regulation of Autonomous Weapons). (2020). *LAWS and export control regimes: Fit for purpose?* iPRAW Working Paper. https://www.ipraw.org/wp-content/uploads/2020/04/iPRAW_WP_ExportControls.pdf
- Kania, E. (2017, June 20). Beyond CFIUS: The strategic challenge of China’s rise in Artificial Intelligence. *Lawfare*. <https://www.lawfareblog.com/beyond-cfius-strategic-challenge-chinas-rise-artificial-intelligence>
- Leung, J., Fischer, S.-C., & Dafoe, A. (2019, August 28). *Export controls in the age of AI. War on the rocks*. <https://warontherocks.com/2019/08/export-controls-in-the-age-of-ai/>
- Metz, C. (2019, January 1). Curbs on A.I. exports? Silicon Valley fears losing its edge. *The New York Times*. <https://www.nytimes.com/2019/01/01/technology/artificial-intelligence-export-restrictions.html>

- OpenAI. (2019). Comments submitted in response to ‘Advance notice of proposed rulemaking: Review of controls for certain emerging technologies’. *Federal Register*, 83(223) (19 Nov. 2018). RIN 0694-AH61.
- Rasser, M., Lamberth, M., Riikonen, A., Guo, C., Horowitz, M., & Scharre, P. (2019, December 17). The American AI Century: A Blueprint for Action. *Center for a New American Security*. <https://www.cnas.org/publications/reports/the-american-ai-century-a-blueprint-for-action>
- Shaw, R. (2016). Export controls and the life sciences: Controversy or opportunity? *EMBO Reports*, 17(4), 474–480.
- Stanley-Lockman, Z. (2019, June 26). Why the Sky is not falling: The diffusion of artificial intelligence. *EurasiaReview*. <https://www.eurasiareview.com/26062019-why-the-sky-is-not-falling-the-diffusion-of-artificial-intelligence-analysis/>
- Stewart, I. J. (2016). *Examining intangible controls: Part I*. Project Alpha, Centre for Science and Security Studies King’s College London.
- Thomsen, R. C., II. (2018). Artificial intelligence and export controls: Conceivable, but counter-productive? *Journal of Internet Law*, 22(5), 15–24.
- TradeSecure. (2019). *Regulating the future*. Blog. <http://tradesecure.net>
- US Department of Commerce, Bureau of Industry and Security. (2020a). Scope of the Export Administration Regulations. In *Export Administration Regulations*. <https://www.bis.doc.gov/index.php/documents/regulations-docs/2382-part-734-scope-of-the-export-administration-regulations-1/file>
- US Department of Commerce, Bureau of Industry and Security. (2020b). Addition of software specially designed to automate the analysis of geospatial imagery to the export control classification number 0Y521 series. *Federal Register*, 85(3), 459–462. <https://www.govinfo.gov/content/pkg/FR-2020-01-06/pdf/2019-27649.pdf>
- US Department of Commerce, Bureau of Industry and Security. (2018). Advance notice of proposed rulemaking: Review of controls for certain emerging technologies. *Federal Register*, 83(223).
- Viski, A., Jones, S., Rand, L., Boyce, T., & Siegel, J. (2020). Artificial intelligence and strategic trade controls. Technical Report. Strategic Trade Research Institute and Center for International and Security Studies at Maryland. <https://strategictraderesearch.org/wp-content/uploads/2020/06/Artificial-Intelligence-and-Strategic-Trade-Controls.pdf>
- Wassenaar Arrangement. (2019). List of dual-use goods and technologies and munitions list, *WA-LIST (19) I*, 5 Dec. 2019. Wassenaar Arrangement Secretariat. <https://www.wassenaar.org/app/uploads/2019/12/WA-DOC-19-PUB-002-Public-Docs-Vol-II-2019-List-of-DU-Goods-and-Technologies-and-Munitions-List-Dec-19.pdf>
- Wassenaar Arrangement. (2011). Elements for objective analysis and advice concerning potentially destabilising accumulations of conventional weapons. As adopted in 1998 and amended by the Plenary in 2004 and 2011. <https://www.wassenaar.org/app/uploads/2019/consolidated/Elements-for-Objective-Analysis.pdf>
- Wassenaar Arrangement. (2003). *Statement of understanding on control of non-listed dual-use items* (Agreed at the 2003 plenary). https://www.wassenaar.org/app/uploads/2019/consolidated/Non-listed_Dual_Use_Items.pdf

Arms Control for Artificial Intelligence



Thomas Reinhold

● آقای هوش مصنوعی ●

🏢 رسانه هوش مصنوعی دانشگاه تهران 🏢

@MrArtificialintelligence

Abstract With military weapon systems getting more and more improved by artificial intelligence and states competing about the leading role in this development, the question arises how arms control measures can be applied to decrease this equipment spiral. The ongoing debates on cyber weapons have already highlighted the problems with controlling or limiting digital technologies, not to mention the dual use problems. While still in an early stage, this chapter develops possible approaches for AI arms control by considering the different life cycle steps of a typical AI enabled system, based on lessons learned from other arms control approaches. It will discuss the different starting points, their arms control potential as well as its limitations to provide a holistic perspective for necessary further develops and debates.

1 Introduction: Or why Hard Arms Control for Artificial Intelligence Should Be Considered

In this book, we look at both the possibility of using artificial intelligence (AI) to foster arms control as well as the dark side—the acceleration of warfare or the possible transfer of decision-making from humans to machines. While AI can foster arms control (see the overview by Schörnig in addition to the individual chapters) at the same time it needs to be controlled. In the debates on cyber arms control and autonomous weapons control, confidence-building measures (CBMs) and increased transparency are often seen as the best outcomes. In some circumstances, as in the case of autonomous weapons, only political declarations remain realistic, but their meaning is unclear until they are actually applied. In short: The arguments why hard, verifiable arms control is not possible are varied and compelling, and they dominate the current discourse.

T. Reinhold (✉)

Chair of Science and Technology for Peace and Security (PEASEC), Department of Computer Science, Technical University of Darmstadt, Darmstadt, Germany
e-mail: reinhold@peasec.tu-darmstadt.de

In this text, however, it will be argued that, at least from a theoretical perspective, there are definitely starting points for quantifiable and verifiable arms control measures in the realm of AI and their realization only needs to be consistently tracked and checked. These approaches are very technology-specific, and it is necessary to unpack the concept of AI to start with. Only when AI is broken down into smaller, manageable and technically relevant parts can promising approaches be identified. This makes it necessary to define AI to begin with.

As many chapters in this book have shown, AI is a well-established concept and includes deterministic variants such as expert systems. The current debates among IT specialists, however, focus on so-called *neural networks* and their recent spin-off variant *deep learning* (see the introductory chapter in this book for details; see also Charniak, 2019; Kersting, 2018). Most civilian applications use this latest form of AI, and it is also widely used in the military realm. Consequently, the structural approach adopted here focuses on neural networks but can be applied analogously to many other forms and variants of AI. To provide a broader perspective and avoid restricting the discussion to a specific technological branch, in the following text, the term *artificial intelligence* will be used, including neural networks and earlier, current or even future forms of machine learning (ML).

As will be seen, the development process or *life cycle* of AI based on ML can be broken down into four components with different but always promising approaches being applied to each individual component.

It goes without saying that this text breaks new ground and, also due to its brevity, only formulates initial thoughts. As a result, no silver bullet can be expected, but ideally a crystallization or starting point at which further discussion can begin. This text does not aim at ending the debate but rather at (re)opening it by introducing technical details in order to overcome the common *it won't work because it's complicated* perspective.

To prove that there are more options for applying arms control measures to AI, the text will present arguments as follows: Section 2 will provide a brief overview of current technological trends and the rise of AI and show why it fosters militarization. Section 3 will briefly examine best practices and established arms control instruments in order to describe the variety of options arms controllers can choose from under varying circumstances. After that, Sect. 4 unpacks the development process of AI and identifies the four key components where arms control measures could start. Section 5 delves deeper into this and identifies the best arms control practices for each of the components. Section 6 addresses the problematic field of verification and how the arms control measures suggested in Sect. 5 could be successfully verified. Section 7 debates potential pitfalls and necessary pre-conditions when such ideas and concepts are applied to real-life AI-enabled weapon systems. Section 8 goes back a step and asks whether CBMs could be a viable alternative to the hard and verifiable measures previously suggested. It concludes that some confidence-building could be done, but also argues that confidence-building alone would not be enough given these options. Finally, Sect. 9 summarizes the text and offers a glimpse into the future.

2 The Rise of Artificial Intelligence and Its Militarization

The technology of ML and more generally AI has taken huge steps in recent years, supported by the miniaturization and performance enhancements of IT. Some use cases that traditionally had been a core application area of AI such as visual pattern recognition have been integrated in broadly distributed consumer electronics in the form of facial recognition, image classification or natural speech analysis and its synthesis (see, for example, the text by Schörnig, 2022). Whereas former AI applications usually focused on one specific task and its optimization, the significantly increased amount of processable data has fostered the development of AI systems that become an integral part of complex applications. Such AI systems process, filter and classify huge amounts of data—partly in real time—and are intended to reduce the data overload of many real-world scenarios for human operators—for example high-frequency trading (Briola et al., 2021), social media hate speech detection (Putri et al., 2020) or automated cybersecurity systems (Belani, 2021). And finally, AI is a core element of the ongoing trend toward autonomous systems that are able to navigate under dynamic, partly uncertain or even unknown environmental conditions (see the text by Dahlmann, 2022). These developments highlight a trend in which the role of integrated AI systems is shifting from being one of many subsystems that deliver input or perform dedicated tasks to becoming the core element that integrates all the different subsystems and generates the final output.

These advances are also affecting military trends, applications and strategic decisions as AI seems to provide the core tool for managing the digitalization of military systems and the necessity to process huge amounts of data into machine- or human-usable information (see the texts by Sauer, 2022; Fischer, 2022). Previous chapters analyzed many of these aspects and discussed the problems and challenges that arise from the application of AI for different military technologies such as the automation of cyber defensive and offensive measures (see the text by Reinhold & Reuter, 2022), robotics and autonomous military vehicles (see the text by Dahlmann, 2022), or even the enhancement and automation of nuclear defense systems (see the texts by Heise, 2022; Baldus, 2022). In addition to direct integration into weapons or weapon control systems, ML algorithms are also being inserted into other military applications such as battlefield management, logistics, recruitment and training of personnel, or other aspects of the complex military administration and bureaucracy (Bundeswehr, 2019).

This development is offering new challenges for the regulation, containment, and non-proliferation of AI as a military technology as well as of AI-enabled military (weapon) systems. As the debates over the militarization of cyberspace have already shown, many established measures of arms control and verification are not applicable to digital technologies because of their specific technical features and thus require new methods (Reinhold & Reuter, 2019a). Whereas political measures such as confidence-building, codes of conduct, or norms debates are already taking place (Paoli et al., 2020), technical approaches that would allow verifiable measures

have not yet been studied. Nevertheless, as Lawrence Lessig once stated: “code is law” (1999), pointing out that software and its underlying code directly reflects the rules and values of its creators who set its capabilities and limits. As any AI is based on code, this is certainly true for the military application of ML. As a work of human beings, it can be controlled in principle, shaped and adjusted to serve a good purpose. Used in the right way, AI can supplement potential international norms restricting the military use of AI with actual control, restriction and verification measures.

The following sections offer preliminary thoughts on how this could be done and what obstacles will be faced.

3 Best Practices and Lessons Learned from Other Technologies

As a first step, it is helpful to look at established arms control measures for other technologies in order to understand the lessons learned and the best practices that can potentially be applied to AI. According to Mölling and Neuneck (2001), the different forms of arms control measures that have been developed for chemical, biological, or nuclear (CBN) weapons as well as conventional forces can be roughly broken down into four groups:

- Declarative measures that are based on agreements of Do’s and Don’ts
- Usage-related measures and regulation
- Trade and proliferation measures
- Information exchange-based measures

Much like the tools developed for the militarization of cyberspace (Reinhold & Reuter, 2019b), digital technologies lack a direct physical representation apart from the interchangeable storage medium required for usage and proliferation-based measures that try to count or track regulated items. The digital information of AI components can be seamlessly copied, cloned, and distributed, which renders any measure that requires physical objects impossible to apply and complicates verification, but favors declaratory, regulatory and information exchange-based approaches. This does not reduce the value of cooperative measures between states or even possible agreements on trade controls of AI components based on company declarations of the traded goods, but it limits the possibilities for controlling compliance with agreements based on objectifiable information or even monitoring other parties to an agreement without their consent or cooperation.

This aspect highlights the necessity of analyzing the technical foundation and characteristics of AI and its components in order to identify features that can be measured and compared. A similar analysis for cyber tools (Reinhold & Reuter, 2019a) concluded that in addition to the technological challenges that have been mentioned, IT-related products actually do provide quantifiable parameters that could be applied for arms control measures. These include:

- The total power supply as well as the current power consumption of IT infrastructures
- The available supply of cooling systems and their thermal power as well as the current heat production of IT infrastructures
- The available network bandwidth capacities as well as the current flowrate of transmitted data via monitored network connections
- The total extent of connections of monitored networks to other external civil or commercial networks (the so-called *peering*) and their maximum possible transmission performance
- The number of staff required for the maintenance of the IT systems

It is thus possible *in principle* to identify measurable and quantifiable aspects of AI where action to implement arms control measures could be taken. As in the case of cyber tools, the following section will unpack the development process or the *life cycle* of AI to identify where in the development process action could be taken and on which key components.

4 The AI Life Cycle: The Components of Artificial Intelligence Applications

The previous chapters in this book and the many approaches used by the authors alone made it clear that there are and have been many different forms of AI and algorithmic approaches. In the current debates, especially the so-called neural networks and their recent spin-off variant *deep learning* (see the introductory chapter in this book for details) play a major role. These approaches, in combination with the processing power of computers and microchips available nowadays, provide the most powerful results and can mimic human intelligence for the first time, as was envisaged in the early years of this field of research (Charniak, 2019). This dominance has led to the fact that neural-network methods are already used synonymously with the term *artificial intelligence* or *machine learning* in many contexts, even when the technological foundations differ (Kersting, 2018). Consequently, the following structural approach focuses on neural networks, although it can be applied analogously to most of the other forms and variants of AI. To provide a broader perspective and avoid restricting the discussion of arms control to a specific technological field, the following text will use the term AI to include neural networks and previous, current or even future forms of ML. Regardless of the different approaches, all AI applications are marked by a specific *life cycle*: from their development to their deployment. This *life cycle* concept reflects the fact that each AI-enabled application passes through different transformation steps that apply initial algorithm and design decisions in technical software components which are then later combined in the final application.

Facilitating a concept from a report on the security of AI (Stiftung Neue Verantwortung, 2019), the following life cycle illustrates these transformation steps for a military AI application:

1. Definition of the goal and the desired capabilities of the AI
2. Acquisition and preparation of the required training data
3. Choosing the required ML methods
4. Learning the implicit input-to-output rules (the so-called *classifiers*) during training with the selected training data
5. Creating a fully trained AI system (the so-called *model*)
6. Deploying the AI into military systems or effectors
7. Applying the military system or effector, probably with a feedback loop and retraining of the model

In these steps the following four different components are always employed in one way or another as part of any AI application and its development:

- A. The training data, that is, the dataset which is given to the algorithm to identify patterns and regularities as well as used for the testing and evaluation of the AI.
- B. The classifiers, that is, the representation of the training goal.
- C. The model, that is, the final data structure which encompasses the learned interrelations and information.
- D. The effector, that is, the actual weapon that achieves the destructive effect under the control of AI.

When debating the chances of implementing arms control measures for AI, it is very important to distinguish systematically between these components, as each of them, together with its associated transformation steps uses different technological approaches and thus provides different technical aspects and characteristics of AI applications that can be used to impose restrictions. This component-centric perspective is useful for maintaining a technical perspective on the possibilities and challenges of AI for arms control. However, while the first three components relate to the development process of the AI algorithm itself, the fourth component is related to its application. As an AI does not directly contain but only controls effectors, it will always be part of a larger military system that provides the actual effectors and must thus be taken into account in arms control measures.

5 The Components of AI Development: Applying Tailored Arms Control Measures

The AI components that have been identified will now be discussed in greater detail, including analysis of the measurable and quantifiable aspects where arms control initiatives can start and where potential technological thresholds between civilian and military AI can be defined. After a review of possible lessons learned from fields



Fig. 1 Data transformation along AI lifecycle development stages. Source: Own illustration

of successful arms control initiatives that might be applied to AI, each subsection that follows will discuss which arms control measures could be applied to each particular component.

5.1 The Training Data

The training data is essential for every AI application that facilitates any kind of learning and adjustment of inner processing capabilities. The data used for training can take many different forms but, in most cases, involves a specific set of information built from streams or batches of raw data and tailored to the specific learning goal as well as the specific variant of learning algorithm. Organization of the data is necessary in order to structure the amount of information presented to an AI algorithm so that it contains enough relevant relationships that can be identified and learned, but does not become too *polluted* with misleading or distracting information. For example, an AI that is required to learn to identify IEDs (improvised explosive devices) in visual information needs to be presented with different images that in the best case contain all different kinds, sizes, shapes, forms of construction, etc., of these devices. A well curated set of training data usually also contains negative data items, in the present example images of devices or objects that are not IEDs. The final curated data set is then usually split into different batches that are used for the training of the AI, for testing the trained model (see Sect. 5.2) with data that has not yet been used for training and which the algorithm has not yet *seen* and a further batch to evaluate the quality of the model. Figure 1 presents an overview of the different stages in the processing pipeline from raw data to applicable training sets as given in an ENISA report (ENISA, 2020).

The different steps in this process can be performed by a single actor or distributed over different institutions or can be provided by commercial vendors or brokers. As the data needs to be collected, processed, and curated in plain text—which means that it cannot be encrypted during this step—it potentially provides options for checking, comparing or verifying against defined principles.¹ This provides the

¹Experience from civilian applications has shown, however, that datasets struggle with unrecognized biases. If, for example, the dataset scarcely features people of color but focuses on white males, the AI might struggle to recognize black faces (Buolamwini & Gebru, 2018).

following access points for control, regulation, or restriction in possible arms control agreements:

- Restrict the use of specific type of information or limit the scope of raw data collections for specific military training goals, for example, for the identification of human combatants.
- Monitor, regulate or restrict the use of specific raw data aggregation infrastructures (such as dedicated cloud services or sensor systems). In particular, the gathering of training data for possible offensive military applications like autonomous weapon systems (AWS) could be restricted to a certain degree to real-world military scenarios such as actual military operations—which is at least impractical—or to dedicated military testing environments designed for such raw data acquisition. The latter could even be passively monitored.
- As far as dedicated vendors, data brokers or curation services are concerned, their commercial activities could be lawfully regulated, in conjunction with appropriate transparency and compliance control measures. This would provide additional measures for proliferation control.
- In addition to the limitation on use off specific raw data, it is also possible to regulate specific kinds of data curation and preparation that reflect specific, limited or unwanted training goals.

5.2 *The Classifiers*

The classifiers of an AI algorithm represent the training goal and in the case of successful training the application quality of the AI system. The exact shape of the classifiers depends strongly on the AI algorithm that is used, but they always reveal the intent of the trainer. Although regulation of this kind of thing is always a challenge for arms control, the following approaches are possible.

- Limit or prohibit the usage of specific types, ranges or characteristics of classifications in order to limit the application scope of AI systems.
- Intentionally limit classifier quality in order to reduce the applicability and degree of autonomy of AI systems and thus enforce closer human interaction and a wider decision range, for example by allowing the classifier to identify humans but not to provide an assessment of their combatant status, leaving this to human judgment.

However, it is not the aim of arms control to check used datasets for biases but to prevent the use of certain datasets which could be used for undesired weapon systems.

5.3 *The Model*

The model of an AI system, as the trained state of an AI that is ready for application, is the embodiment of the intended goals implicit to the training data and the selected classifiers. With regard to the AI system itself, the model is the final product that could be built into the designated external system which it supports or controls. Whereas some models could be highly specific for a designated use case and external system, others could be more *off the shelf*, generalized and applicable to a huge variety of external applications. Thus, the following possible arms control measures exist:

- Regulation or restriction of the proliferation of models trained for specific military purposes such as distinguishing between civilians and human combatants.
- Control or prohibition of trade in models in conjunction with Wassenaar-like information and transparency measures.
- Restriction of the use of specifically trained models, either for direct application in an external system or for use as the basis of further AI training scenarios.

The ongoing trend to miniaturization and specialization of microchips that provide among other things AI-optimized hardware also requires regulation. But since such hardware parts are not designed for a specific use case but rather to be equipped with a dedicated AI model, their regulation raises strong dual-use concerns, as will be discussed below.

5.4 *The Effectors*

An AI-enabled application has—in contrast with most other militarily weaponized technology—no direct effect on its environment. Instead, the AI will always be part of larger weapon systems in which it controls specific aspects of the system or controls it completely up to the release of the actual effector—tasks that mostly had been or are still assigned to human operators. In many cases the weapon system itself is not a new development and the AI is simply an extension or upgrade, enhancing systems like air defense, uncrewed vehicles, battlefield command and control, or cyber defense measures. This means that in the best-case scenario, these weapon systems are already part of arms control agreements that can be adapted to include AI-specific regulations. A second aspect of this relationship is that it directly relates to the question of the limitations and boundaries of the autonomy of weapon systems or trigger decisions and the debates about control of the acceptable extent of such capabilities. In conclusion, the following arms control measures are applicable:

- Extend existing arms control treaties to include the enhancement or replacement of components of the regulated items and technology with AI applications or include these aspects in negotiations on the renewal of terminated treaties.

- Include AI applications or systems that are intended to be integrated into weapon systems in existing arms trade and non-proliferation agreements such as the Wassenaar Arrangement (WA).
- Expand discussions, international security debates, and treaty negotiations on the regulation of AWS to encompass various aspects of and potential integration of AI and its consequences—including its regulation according to the International Humanitarian Law (IHL).

6 Verification

The previous section has shown that there are indeed starting points for applying hard arms control measures to AI if the dazzling term *AI* is broken down into technically elementary components. However, probably the hardest part of applying arms control measures for digital goods is the challenge of how compliance with agreements can be verified (see the text by Schörnig, 2022). This also applies to AI algorithms and the situation is additionally complicated by the black box character of an AI (see the text by Verbruggen, 2022). This technical aspect arises from the fact that the model of an AI does not provide a human-readable or comprehensible representation of the learned states and the algorithmic micro-decisions it makes. For current AI algorithms, all that can be seen is the output arising from a given input, not the path that led to this conclusion. This raises the question of which parts of an AI application could be controlled in terms of defined thresholds or prohibitions. The following list presents initial ideas for dealing with this challenge. It is not meant to be complete and is highly dynamic in view of emerging technical developments in the field of AI.

- Training a clean model with the data that was allegedly used for the original AI must create a model that works identically to and generates the same results as the defined set of testing input. This makes it possible to verify whether an AI has been trained with a set of training data that complies with agree-upon rules. This method is limited to static AI applications that are not re-trained or otherwise adapted during their real-life application, as adaptation changes their internal state and thus undermines comparability.
- To verify that decisions made by an AI comply with certain rules, it is possible to use a set of test data specifically constructed to contain triggering input which will lead to a specific output. As a trivial example, an AI could be trained to identify tanks in images and tag them as military targets under the restriction that it will not tag other objects or even humans as targets and will untag tanks that are relatively close to humans. A test set of images would include images of tanks as well as humans in different surroundings and combinations. Tested against these images, the AI must only tag the tanks that are not surrounded by humans.
- Newer technological developments of specific AI algorithms may provide the technical means for the verification of decisions. A research trend involving

so-called *explainable AI* (Vilone & Longo, 2020) provides a retraceable input-to-output path which at least makes it possible to understand the technical process that resulted in a particular decision and permits conclusions on the influence of the training on the real-world performance that followed it. Even if this procedure is not capable of identifying the effects of specific training input, it can provide understanding of how specific clusters of training data modified the final model and its data processing. Since such algorithms require the storage of additional information as well as the necessary data processing, such features usually reduce the overall performance of the AI algorithm. As it would unravel the black box character as an important precondition for arms control, explainable AI can provide an important tool, but will have to be made mandatory in agreements in order to be implemented.

- A final challenge lies in the task of verifying whether an AI has been used as part of an existing system without deep analysis of the operation system. In most cases this is not accepted by treaty members. This resembles the often-cited and still valid idea of the Turing test in connection with the choice between an AI and some other form of deterministic algorithm with hard-wired instructions. Under optimal conditions the latter will always provide an output that is predictable, as it can be calculated externally as long as the hard-wired instructions are known. An AI on the other hand is designed to provide the best possible approximations to the exact result for an input that has not been used during the training phase. These differences between the actual and the exact result might be used to identify the application of an AI.

In the military there is a saying *it takes one to get one* meaning that in certain situations symmetry is the only possible response or is needed to counter a specific capability. This poses the interesting challenge of using an AI to verify other AI. As the algorithms involved in ML are—in addition to other uses—perfectly applicable to detecting patterns within unknown data or separating and classifying complex information, it is at least theoretically possible to train a *verification AI* with the output from another AI that needs to be monitored. The results yielded by the *verification AI* could then make it possible to draw conclusions concerning the learned processing rules of the AI being checked or the training input it is assumed to have received. Even if such thinking is futuristic at present and applicable measures have yet to be developed, it could serve as the basis for establishing measures for controlling compliance with agreed rules.

7 Pre-conditions and Pitfalls for Arms Control

Many of the ideas discussed and considered above are currently still no more than theory and appropriate technical approaches need to be developed, tested, and—conceivably if ever—installed as measures for arms control. This is, on the one hand, a direct result of the fact that AI is a relatively new topic in military technology,

whose capabilities have been boosted by the development of more efficient algorithms alongside dedicated hardware. On the other hand, its implications and limitations are not as yet fully understood and arms control measures for AI technology must consider the specific conditions discussed above.

The first of these is the obvious and presumably most influential issue of the highly dual-use character of AI and ML. As AI and its components are inherently only parts that are included in more comprehensive systems for specific tasks, the regulation of explicitly military AI will turn out to be inefficient. Although AI applications that are specifically trained with military-grade information and intended to cover specifically military use cases presumably either already exist or will eventually do so, in most cases more generic AI components will be produced and acquire capacities (such as image recognition, information clustering, etc.) that will later only need to be adapted to specific tasks. This aspect also relates to AI-specific hardware that is experiencing strong demand and a corresponding driving force in civil commercial products such as consumer electronics. The further miniaturization of such generically applicable technology will probably further strengthen a trend toward cheap off-the-shelf hardware that is ready to be deployed.

A further aspect relates to the current technological imbalance of AI technology. Although a great deal of groundwork has been carried out in recent decades and published in scientific journals and conference proceedings, the current trend in the implementation of AI in real products is being driven by a small number of technological global players that hold the intellectual property rights. It is foreseeable that these companies, and with them the states where they operate, will try to defend this head start in order to preserve the advantages they have gained from this technology in both commercial and also military domains. This imbalance between the *haves* and the *have-nots* will probably complicate the establishment of arms control measures as it has to deal with inherently opposing interests. In addition, AI research and its development have a strong dual-use character. As the actual use of an AI is primarily determined by its training, the underlying algorithms involved in how exactly the model is developed on the basis of input information or how classifiers are created and applied is the same for military as well as civil uses and application. This aspect also includes dedicated AI hardware such as specific microchips that are optimized to perform the required AI calculations or feature a specific technical design that is adjusted to AI models such as neural networks. This complicates the regulation of AI algorithms and their implementation in specific hardware.

Another issue relates to the problems that have already been discussed regarding the technical challenges involved in verifying AI arms control measures. The characteristics of digital goods provide many chances and opportunities for hiding non-compliant behavior while simultaneously hindering effective control mechanisms. In addition, the availability of related commercial products makes it easier to establish a dedicated domestic industry for military-grade AI. This might either prevent states from joining such *toothless* agreements or—on the contrary—might even offer states an incentive to dishonestly sign treaties safe in the knowledge that

non-compliance is not trackable. This challenge may be eased with further technological developments but so far is a game stopper.

The final aspect that will probably hinder the establishment of arms control measures for AI concerns is the perception of this technology as mostly unproblematic and not dangerous enough. In most proposals, research projects or statements from military decision-makers, AI is seen as an enabler for military systems or as an enhancement for human tasks. Although debates in other areas such as lethal autonomous weapon systems (LAWS) discuss the threats and problems that arise from decisions made autonomously by machines, these concerns have so far not been included in AI debates to a sufficient degree. As long as AI is not perceived as another aspect of the same problem, there will not be sufficient incentives for states to debate its regulation and the limitation of its military application.

8 Confidence-Building Measures for Military AI Applications: An Alternative?

The preceding sections have shown that the application of *hard* and verifiable arms control measures is not impossible. But just starting to think about the possibilities requires extensive technical knowledge—knowledge that arms control experts often do not possess. Consequently, the first step toward actual arms control agreements has often been the establishment of confidence via CBMs. In most cases this step has involved, among other things, the exchange of information about national security interests and concerns about shifting military capabilities resulting from technological developments as well as technical details of new developments. These measures for achieving transparency are intended to allow potential adversaries to gain an understanding of the military impact of the adoption of new technological developments as well as of their limitations. With regard to the influence of AI on military developments, the following details of the different components of an AI could be made available as part of CBMs in order to understand its impact:

- Samples of the training data related to the intended capability of the AI
- Training environments or data aggregations sources
- The classifiers and the features that are intended to be detected and processed for the output of the AI
- Details of the application of the AI and the facilitation of its output with regard to the complexity and the degree of freedom that the AI's decisions are used for
- Details of the system that the AI is part of (e.g., effectors, military relevance, and facilitation)
- Information on the structural changes in tactics or on organizational changes where AIs are used to enhance human decisions or replace them

Regarding the similarities that AI shares with other digital technologies, it is important to highlight the contrasting conceptualizations of AI in existing debates on CBMs for cybersecurity and cyberspace in international forums like the Organization for Economic Co-operation and Development (OECD) or the United Nations (UN). As military capabilities are mostly shaped by human skills in cooperation with intelligence-gathering operations, the debates on cyber CBMs mostly cover its impact on military defensive and offensive strategies, but seldom involve technical details or technological knowledge. On the other hand, the military development and application of AI is driven far more strongly by active scientific research on AI algorithms and dedicated hardware and is thus influenced by issues of intellectual property and maintaining a technological edge in knowledge. Thus, although it might be meaningful to promote existing cyber forums on CBMs, these debates will probably face greater reluctance by participating parties to share the technical details mentioned above and may have to focus more on strategic goals.

9 Conclusion: Or How AI May Develop and What Arms Control Can Do About It

When looking at current trends in AI, it is safe to conclude that one way or another AI will find its way into military applications. Even if the current level of attention is reduced or has to face the inherent limitations of this technology, the normative power of the factual as well as the money currently being spent on Research and Development (R&D) will bring the world AI-enabled military systems. This will probably happen regardless of whether they actually perform better, as long as they promise to shorten the sensor-to-trigger loop or otherwise seem to supersede human cognition and reaction limitations. On the other hand, it is doubtful that we will see any kind of an envisage complex AI systems that integrates and controls complex battlefield activities in the near future because the complexity of such activities conflicts with the single-purpose performance of AI algorithms. At best, there will be an integration of multiple specific AI applications, each optimized and facilitated for a dedicated task that will be integrated into such systems, much as is already the case for self-driven cars that consist of multiple interoperating AI applications. Another issue is the currently strongly divided technology ownership. It is quite possible that, regardless of its actual usage, the most advanced AI countries will continue to perfect AI capabilities or even further extend them in order to maintain their current advantage. This could result in strategic benefit or be at least a bargaining chip in international power struggles. In addition, as AI is—in contrast, to for example the cyberspace area—strongly connected with intellectual property rights and technological research and knowledge, this will probably be closely accompanied by economic and trade restrictions. As AI hardware becomes more and more important and a question of performance, such issues could even spill over to the current international disputes and struggles to create national sovereignty over microchip

design and production (Kleinhans & Baisakova, 2020). From the standpoint of military technology, the ongoing trend toward miniaturization of computation devices that also includes AI hardware may foster and accelerate a shift from current military R&D projects involving large monolithic AI systems for complex tasks to the integration of dedicated AI capabilities into small military systems and consumables such as small arms, land mines and ammunition. As small arms are still the real weapon of mass destruction (WMD), AI-enabled small arms with self-guiding ammunition might be even more terrifying and deadly.

This leaves a great deal of work for further arms control approaches and requires substantial convincing of national and international actors. It probably also means that in the near future AI will become one of the many factors that need to be discussed and considered in connection with many existing weapon systems and military capabilities. This could also increase the necessity of including AI in existing arms control treaties. Measures for AI face issues similar to those involved in the militarization of cyberspace, where many established arms control approaches have not worked and have thus led to a need for new technical solutions and tools for verification. As AI and cyberspace share a great deal of underlying technology it probably makes sense to combine discussions and the development of arms control tools based on these technologies. On the other hand, AI-enabled applications or military systems will still rely on small-scale single-problem AI solutions so that there will still be opportunity for approaches to its regulation that focus on specific details, technical features, or capabilities, without the necessity of tackling the sci-fi vision of a *super-AI*. This also means that verification measures—despite the problems mentioned—could be built upon very detailed features, which, from a technical perspective, leaves room for optimism. And that is something that arms control has always needed.

References

- Baldus, J. (2022). Doomsday machines? Nukes, nuclear verification and artificial intelligence. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.
- Belani, G. (2021). The use of artificial intelligence in cybersecurity: A review. *IEEE Computer Society*. <https://www.computer.org/publications/tech-news/trends/the-use-of-artificial-intelligence-in-cybersecurity>
- Briola, A., Turiel, J., Marcaccioli, R., & Aste, T. (2021). Deep reinforcement learning for active high frequency trading. *arXiv*. <https://doi.org/10.48550/arXiv.2101.07107>
- Bundeswehr. (2019). *Künstliche Intelligenz in den Landstreitkräften*. <https://www.bundeswehr.de/de/organisation/heer/aktuelles/kuenstliche-intelligenz-in-den-landstreitkraeften-156226>
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research, 81* (pp. 1–15). 2018 Conference on fairness, accountability, and transparency.
- Charniak, E. (2019). *Introduction to deep learning*. MIT. <https://doi.org/10.5555/3351847>
- Dahlmann, A. (2022). Armament, arms control and artificial intelligence: The impact of software, machine learning and artificial intelligence on armament and arms control. In T. Reinhold &

- N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.
- ENISA. (2020). *Artificial Intelligence Cybersecurity Challenges - Threat Landscape for Artificial Intelligence*. <https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges>
- Fischer, S.-C. (2022). Military AI applications: A cross-country comparison of emerging capabilities. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.
- Heise, A. (2022). AI, WMD and arms control III: The case of nuclear testing. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.
- Kersting, K. (2018). Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines *Frontiers in Big Data, 1*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7931929/>
- Kleinhans, J.-P., & Baisakova, N. (2020). *The global semiconductor value chain: A technology primer for policy makers*. Stiftung Neue Verantwortung. https://www.stiftung-nv.de/sites/default/files/the_global_semiconductor_value_chain.pdf
- Lessig, L. (1999). *Code and other laws of cyberspace*. Basic Books, Inc.
- Mölling, C., & Neuneck, G. (2001). Präventive Rüstungskontrolle und Information Warfare. In *Rüstungskontrolle im Cyberspace. Perspekt. der Friedenspolitik im Zeitalter von Comput. Dokumentation einer Int. Konf. der Heinrich-Böll-Stiftung am 29./30. Juni 2001 Berlin* (pp. 47–53).
- Persi Paoli, G., Vignard, K., Danks, D., & Meyer, P. (2020). *Modernizing arms control: Exploring responses to the use of AI in military decision-making*. UNIDIR. <https://unidir.org/publication/modernizing-arms-control>
- Putri, T. T. A., Sriadhi, S., Sari, R. D., Rahmadani, R., & Hutahaean, H. D. (2020). A comparison of classification algorithms for hate speech detection. *IOP Conference Series: Materials Science and Engineering, 830*(3). <https://iopscience.iop.org/volume/1757-899X/830>
- Reinhold, T., & Reuter, C. (2019a). Arms control and its applicability to cyberspace. In C. Reuter (Ed.), *Information Technology for Peace and Security - IT-applications and infrastructures in conflicts, crises, war, and peace* (pp. 207–231). Springer Fachmedien Wiesbaden.
- Reinhold, T., & Reuter, C. (2019b). Verification in cyberspace. In C. Reuter (Ed.), *Information Technology for Peace and Security - IT-applications and infrastructures in conflicts, crises, war, and peace* (pp. 257–275). Springer Fachmedien Wiesbaden.
- Reinhold, T., & Reuter, C. (2022). Cyber weapons and Artificial Intelligence – Impact, influence and the challenges for arms control. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.
- Sauer, F. (2022). The military rationale for AI. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.
- Schörnig, N. (2022). Introduction. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.
- Stiftung Neue Verantwortung. (2019). *Securing artificial intelligence*. https://www.stiftung-nv.de/sites/default/files/securing_artificial_intelligence.pdf
- Verbruggen, M. (2022). No, not that verification: Challenges posed by testing, evaluation, validation and verification of artificial intelligence in weapon systems. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.
- Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: A systematic review. *arXiv*. <https://doi.org/10.48550/arXiv.2006.00093>

● آقای هوش مصنوعی ●

🏢 رسانه هوش مصنوعی دانشگاه تهران 🏢

@MrArtificialintelligence