Routledge Studies in Conflict, Security and Technology

ARTIFICIAL INTELLIGENCE AND INTERNATIONAL CONFLICT IN CYBERSPACE

Edited by
Fabio Cristiano, Dennis Broeders, François Delerue,
Frédérick Douzet, and Aude Géry

Artificial Intelligence and International Conflict in Cyberspace

This edited volume explores how artificial intelligence (AI) is transforming international conflict in cyberspace.

Over the past three decades, cyberspace developed into a crucial frontier and issue of international conflict. However, scholarly work on the relationship between AI and conflict in cyberspace has been produced along somewhat rigid disciplinary boundaries and an even more rigid sociotechnical divide – wherein technical and social scholarship are seldomly brought into a conversation. This is the first volume to address these themes through a comprehensive and cross-disciplinary approach. With the intent of exploring the question 'what is at stake with the use of automation in international conflict in cyberspace through AI?', the chapters in the volume focus on three broad themes, namely (1) technical and operational, (2) strategic and geopolitical and (3) normative and legal. These also constitute the three parts in which the chapters of this volume are organised, although these thematic sections should not be considered as an analytical or a disciplinary demarcation.

This book will be of much interest to students of cyber-conflict, AI, security studies and International Relations.

Fabio Cristiano is an Assistant Professor of Conflict Studies at Utrecht University, where he teaches in the MA in Conflict Studies & Human Rights and the Minor in Conflict Studies. He is an Associate Fellow of The Hague Program on International Cyber Security at Leiden University and holds a PhD in Political Science from Lund University.

Dennis Broeders is a Full Professor of Global Security and Technology at the Institute of Security and Global Affairs (ISGA) of Leiden University, the Netherlands. He is the Senior Fellow of The Hague Program on International Cyber Security and project coordinator at the EU Cyber Direct Program.

François Delerue is an Assistant Professor of Law and a member of the Jean Monnet Centre of Excellence for Law and Automation (Lawtomation) at IE University. His book *Cyber Operations and International Law* was published in 2020.

Frédérick Douzet is a Professor of Geopolitics at the University of Paris 8, Director of the French Institute of Geopolitics research team (IFG Lab) and Director of the Center Geopolitics of the Datasphere (GEODE). She has been a senior member of the Institut Universitaire de France since 2022 and a member of the French Defense Ethics Committee since 2020.

Aude Géry is a Post-doctoral Fellow at GEODE. She also co-chaired the Committee on Digital Challenges for International Law set up for the 150-year anniversary of the International Law Association.

Routledge Studies in Conflict, Security and Technology
Series Editors: Mark Lacy, Lancaster University, Dan Prince, Lancaster University, and Sean Lawson, University of Utah

The *Routledge Studies in Conflict, Technology and Security* series aims to publish challenging studies that map the terrain of technology and security from a range of disciplinary perspectives, offering critical perspectives on the issues that concern publics, business and policymakers in a time of rapid and disruptive technological change.

Emerging Security Technologies and EU Governance
Actors, Practices and Processes
Edited by Antonio Calcara, Raluca Csernatoni and Chantal Lavallée

Cyber-Security Education
Principles and Policies
Edited by Greg Austin

Emerging Technologies and International Security
Machines, the State and War
Edited by Reuben Steff, Joe Burton and Simona R. Soare

Militarising Artificial Intelligence
Theory, Technology and Regulation
Nik Hynek and Anzhelika Solovyeva

Understanding the Military Design Movement
War, Change and Innovation
Ben Zweibelson

Artificial Intelligence and International Conflict in Cyberspace
Edited by Fabio Cristiano, Dennis Broeders, François Delerue, Frédérick Douzet, and Aude Géry

For more information about this series, please visit: https://www.routledge.com/Routledge-Studies-in-Conflict-Security-and-Technology/book-series/CST

Artificial Intelligence and International Conflict in Cyberspace

Edited by
Fabio Cristiano, Dennis Broeders,
François Delerue, Frédérick Douzet,
and Aude Géry

Routledge
Taylor & Francis Group
London and New York

First published 2023
by Routledge
4 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
Names: Cristiano, Fabio, editor. | Broeders, D. (Dennis), editor. | Delerue, François, 1987– editor. | Douzet, Frédérick, editor. | Géry, Aude, editor.
Title: Artificial intelligence and international conflict in cyberspace / edited by Fabio Cristiano, Dennis Broeders, François Delerue, Frédérick Douzet, and Aude Géry.
Description: Abingdon, Oxon ; New York, NY : Routledge, 2023. | Includes bibliographical references.
Identifiers: LCCN 2022060760 (print) | LCCN 2022060761 (ebook) | ISBN 9781032255798 (hardback) | ISBN 9781032255873 (paperback) | ISBN 9781003284093 (ebook)
Subjects: LCSH: Cyberspace operations (Military science) | Artificial intelligence—Political aspects. | Computer security—Political aspects. | National security.
Classification: LCC U167.5.C92 A78 2023 (print) | LCC U167.5.C92 (ebook) | DDC 355.4—dc23/eng/20230417
LC record available at https://lccn.loc.gov/2022060760
LC ebook record available at https://lccn.loc.gov/2022060761

ISBN: 978-1-032-25579-8 (hbk)
ISBN: 978-1-032-25587-3 (pbk)
ISBN: 978-1-003-28409-3 (ebk)

DOI: 10.4324/9781003284093

Typeset in Bembo
by codeMantra

Contents

Contributors

Alexander Blanchard is the Dstl Digital Ethics Research Fellow at The Alan Turing Institute, London. His research is on the ethics of political violence and emerging technology.

Dennis Broeders is a Full Professor of Global Security and Technology at the Institute of Security and Global Affairs (ISGA) of Leiden University, the Netherlands. He is the Senior Fellow of The Hague Program on International Cyber Security and project coordinator at the EU Cyber Direct Program.

Fabio Cristiano is an Assistant Professor of Conflict Studies at Utrecht University, where he teaches in the MA in Conflict Studies & Human Rights and the Minor in Conflict Studies. He is an Associate Fellow of The Hague Program on International Cyber Security at Leiden University and holds a PhD in Political Science from Lund University.

François Delerue is an Assistant Professor of Law and a member of the Jean Monnet Centre of Excellence for Law and Automation (Lawtomation) at IE University. His book *Cyber Operations and International Law* was published by Cambridge University Press in 2020 and was awarded the 2021 Book Prize of the European Society for International Law.

Frédérick Douzet is a Professor of Geopolitics at the University of Paris 8, Director of the French Institute of Geopolitics research team (IFG Lab) and Director of the Center Geopolitics of the Datasphere (GEODE). She is a senior member of the Institut Universitaire de France since 2022 and a member of the French Defense Ethics Committee since January 2020.

Andrew C. Dwyer is a Lecturer (Assistant Professor) in Information Security at Royal Holloway, University of London and leads the Offensive Cyber Working Group. His research ranges from malware, offensive cyber capabilities, to models and environments in cybersecurity, building on his DPhil (PhD) at the University of Oxford in 2019.

Elizabeth Edgar is currently an independent consultant and advisor. She was Senior Principal Psychologist at the UK Defence Science Technology Laboratory, and Assistant Director of Research for the British Army.

Aude Géry is a Post-doctoral Fellow at GEODE. Her research focuses on the international regulation of the digital space. Her thesis, 'International law and the proliferation of offensive cyber capabilities', was awarded several prizes, including from the French branch of the International Law Association. She co-chaired the Committee on *Digital Challenges for International Law* set up for the 150-year anniversary of the International Law Association.

Kacper T. Gradon, PhD, DSc, is an Associate Professor in the Department of Cybersecurity, Warsaw University of Technology, Honorary Senior Research Fellow at University College London (Department of Security and Crime Science) and Faculty Affiliate and Visiting Fulbright Professor at University of Colorado Boulder (Institute of Behavioral Science, Prevention Science Program), WHO-accredited Global Infodemic Manager, double TED Speaker and former Criminal Intelligence professional.

Jeppe T. Jacobsen is an Assistant Professor at the Institute for Military Technology, the Royal Danish Defence College where he focuses his research on new military technologies and states' political and military behaviour in cyberspace. Jeppe is the editor of the *Scandinavian Journal of Military Studies*.

Jack Kenny is a Research Fellow in International Law at the British Institute of International and Comparative Law and a Research Fellow at the Hebrew University of Jerusalem. He obtained his doctorate from the University of Oxford where his research focused on the relationship between the principle of sovereignty and state cyber operations.

Tobias Liebetrau is a Researcher at the Centre for Military Studies at the University of Copenhagen. His research explores international political aspects of cybersecurity, digital technology and infrastructure.

David McNeish is a Human Factors Engineer at the UK Defence Science Technology Laboratory (Dstl). David was previously Senior Technology Advisor at the Centre for Data Ethics and Innovation (CDEI), part of the UK Government Department for Digital, Culture, Media & Sport.

Wesley R. Moy, PhD, is an Adjunct Lecturer in the Global Security Studies Program in the Advanced Academic Program at Johns Hopkins University. He was also an Adjunct Professor at the National Intelligence University in the U.S. Office of the Director of National Intelligence, a retired intelligence officer from the U.S. Department of Homeland Security and retired U.S. Army strategic intelligence officer. He is a plank holder of the National Counterterrorism Center and has military command experience at the company, battalion and brigade levels. His expertise is in homeland security intelligence and counterterrorism.

Louis Perez is a PhD candidate at the Université Paris Panthéon-Assas (Centre Thucydide) and a teaching fellow at the Université Paris Nanterre. His doctoral research benefits from a grant of the French Ministry of Defence and focuses on International Law and Military Applications of Artificial Intelligence.

Simona R. Soare is a Research Fellow with the International Institute for Strategic Studies (IISS). Simona specialises in defence innovation and emerging and disruptive technologies with a focus on North American and European defence, NATO and the EU. Simona holds a PhD in International Security (2011) and is a US Department of State Fellow as well as a Denton Fellow.

Arun Mohan Sukumar is a Post-doctoral Research Fellow at The Hague Program on International Cyber Security, Leiden University. He is a co-editor of *Multistakeholder Diplomacy: Building an International Cybersecurity Regime* (Edward Elgar Publishing, forthcoming) and the author of *Midnight's Machines: A Political History of Technology in India* (Penguin Random House India, 2019).

Mariarosaria Taddeo is an Associate Professor and Senior Research Fellow at the Oxford Internet Institute, University of Oxford, where she is Deputy Director of the Digital Ethics Lab, and is Dstl Ethics Fellow at the Alan Turing Institute, London.

Acknowledgements

This edited volume came out of a conference co-organised in September 2021 by The Hague Program on International Cyber Security of Leiden University and the Géode Center of the University of Paris 8. The editors would like to thank Corianne Oosterbaan and Alix Desforges for their help with the organisation of the conference and all the other supporting work that went into making this book. The Hague Program on International Cyber Security gratefully acknowledges the generous financial support of the Netherlands' Ministry of Foreign Affairs for the program and this publication.

1 Artificial intelligence and international conflict in cyberspace

Exploring three sets of issues

Fabio Cristiano, Dennis Broeders, François Delerue, Frédérick Douzet and Aude Géry

Introduction

Over the last three decades, cyberspace developed into a crucial frontier and issue of international conflict. Disproving the initial fear-mongering expectations of fully-fledged wars occurring in and through cyberspace, this conflict increasingly unfolds 'away from' the traditional categories and thresholds of war and peace.[1] As argued by Lucas Kello, cyberspace is neither truly at war nor at peace but maintains a constant condition of 'unpeace.'[2] International conflict in cyberspace primarily occurs in the so-called grey zone and often pertains to the domain of information, data, and their manipulation, culminating in acts of espionage, sabotage, and subversion. As empirical evidence overwhelmingly shows, confrontation in cyberspace mostly consists of low-impact hacking, espionage, disinformation, and surveillance.[3] In light of this, recent scholarly work interrogates whether we should consider conflict in cyberspace as an 'intelligence competition' rather than through the lenses of traditional warfare.[4] At the same time, this does not mean that we should think of cyberspace as the peaceful, yet ungoverned and ungovernable, oasis envisioned by cyber libertarians in the early days of the internet[5] – quite the opposite. States now conventionally conceive of cyberspace as an issue of national security and increasingly safeguard and promote their national interests through both defensive strategies and offensive operations in cyberspace.[6]

In a context where data and information have become increasingly important, it comes as no surprise that the development and application of artificial intelligence (AI) have gained momentum in the various discourses about international conflict in cyberspace. AI technologies – such as machine learning, natural language processing, quantum computing, neural networks, and deep learning – provide military and intelligence agencies with new operational solutions for predicting and countering threats as well as for conducting offensive operations in cyberspace. Besides automating the production of knowledge about cyber threats, AI can also automate decision-making, which could 'dilute' the role of (human) political agency as an element of international conflict in cyberspace. Concerns at the core of the international

DOI: 10.4324/9781003284093-1

debate about Lethal Autonomous Weapon Systems (LAWS) would then also enter the debate about cyber conflict. Moreover, the operational entanglement of AI technologies in cyberspace further blurs the already contested lines between defence and offence in cyberspace,[7] while also challenging the divide between cyber conflict and information operations.[8] Besides opening up new operational milieus, the adoption of AI-enhanced cyber capabilities also represents an important strategic asset for states, with the ongoing global race towards the adoption of these technologies fully embedded in broader geopolitical conflicts, deterrence, securitisation strategies, and techno-nationalist narratives, such as those about digital sovereignty.[9]

The entanglement of AI technologies with cyber conflict raises several issues primarily related to human-machine interaction, the role of (big) data in society, great powers competition, and regulation. While creating the 'illusion' of scientific and data-driven security, delegating security functions to independent machines might expose networks to a whole variety of new risks emerging because of autonomy and automation.[10] Potential biases in the mechanical processing of data can lead to miscalculations and the creation of a broader 'attack surface' and vulnerability for the systems that AI purports to protect. Similarly, the global race towards the acquisition of these technologies also risks further intensifying and polarising international conflict in cyberspace.[11] For these reasons, AI technologies have also gained interest as a normative issue across ethical and legal debates on responsible (state) behaviour in cyberspace – although the debate about autonomy has not fully crossed over from the military domain 'proper' to that of cyber conflict yet.[12] As this volume shows, specific regulatory frameworks and legislations might be required to capture AI as both a potential asset and threat to national security and to the 'open and secure' cyberspace that some countries seek to uphold.

With the intent of exploring the question 'what is at stake with the use of automation in international conflict in cyberspace through AI?', this volume focuses on three themes, namely: (1) technical and operational, (2) strategic and geopolitical, and (3) normative and legal. These also constitute the three parts in which the chapters of this volume are organised. Scholarly work on the relationship between AI and conflict in cyberspace has been produced along somewhat rigid disciplinary boundaries and an even more rigid sociotechnical divide – wherein technical and social scholarship are seldomly brought into a conversation. This volume addresses these themes through a comprehensive and cross-disciplinary approach. In this sense, the organisation of the volume in three parts should not be considered as an analytical or, even less so, a disciplinary demarcation. The remainder of this introductory chapter outlines, and provides context for, the main debates of each of the three parts of the volume.

Technical and operational considerations

AI has emerged as the defining technology of our times and seems to epitomise the ultimate innovation that everybody wants and about which

everybody is 'concerned.' States often have a techno-optimistic view of new technologies and look favourably at the prospect of rationalising and perfecting governance through automation,[13] with AI currently being applied to wide and diverse governance domains and issues. The allure of the concept of 'AI' is perhaps best caught by the fact that many applications in government (and outside of it) would still be more aptly labelled as 'classic' automation rather than AI or the introduction of autonomy in systems. However, developments in AI do start to permeate traditional governance by expanding the range, scale, and complexity of operations that can be meaningfully automated, including those associated with cybersecurity. When compared to other governance branches, the application of AI technologies in cybersecurity represents however less of an innovation. Already in the 1990s, machine learning and neural networks were, for instance, applied to the filtering and classification of spam emails.[14] After all, automation constitutes an inherent feature of internet technology and computation. What is relatively new, and of main interest for this volume, is the internationalisation and 'datification' of conflict in cyberspace, where the potential of AI marks a new operational phase through autonomy.

From an operational perspective, AI technologies promise to contribute to one of the core dynamics of international conflict in cyberspace: the identification of vulnerabilities through timely and effective interpretation of data – for either defence or offence. That is, AI has the potential to make conflict in cyberspace more knowable and predictable. When considering aspects of automation and machine autonomy in the context of international conflict in cyberspace, the ability of intelligent machines to make operational choices – at different degrees of independence – points to the question of *who* the actual enactors of international conflict in cyberspace are. As will be further discussed in the third part of this volume, this question is not only analytical or technical. Knowing who enacts conflict in cyberspace also intimately pertains to questions of responsibility.[15] In a context where agency appears to be already diluted through networks, and socio-technical assemblages, exploring the AI-cyber nexus primary means to explore human-mechanic interactions.[16]

The question of autonomy and AI raises an operational interrogative related to the 'place' of humans in relation to the so-called 'loop' of operational decision-making. This dilemma has been foremostly articulated in debates about LAWS where the central question remains whether humans shall be placed in, on, or outside of this loop.[17] In Chapter 2 of this volume, Andrew Dwyer directly addresses this question by analysing the role of deep reinforcement learning (RL) algorithms to question assumptions that AI technologies make conflict in cyberspace more knowable. It argues that, by recognising, performing, and transforming the who, where, and how, of international conflict in cyberspace, AI constitutes more than an epistemic tool for improving operations. In this sense, the chapter also complicates normative considerations about controllable and ethically accountable AI systems and about the place of the human 'in' the loop.

One of the core technical promises of AI for cybersecurity consists in what Tim Stevens defines as a shift 'from known threats to the prolepsis of as-yet-unknown threats and into an anticipatory posture that has received much attention in the critical security literature.'[18] In Chapter 3, Wesley Moy and Kacper Gradon explore the various potential applications of AI in the propagation of disinformation and misinformation, as well as in the context of hybrid and asymmetric warfare. By analysing two methodologies – namely 'Generative Adversarial Networks' and 'Large Language Models' – this chapter explains the relevance of AI for understanding how links are formed, how information is disseminated, and how information can influence opinions and actions in social networks. Taken together, the contributions to the first part of the volume indicate that, while enhancing operational efficiency, AI applications do not necessarily 'make' international conflict more known/predictable and cybersecurity more human-centric. Rather, autonomy and automation further contribute to the problematic understanding of cyberspace as a primarily technical and operational issue or domain.

Strategic and geopolitical considerations

Looking beyond its technical possibilities and operational dilemmas, AI is set to become a constitutional component of economic, political, and military power in the digital age. With the return of great-power competition and the constant contestation and confrontation between states in cyberspace, AI is undergoing a process of securitisation that transforms this dual technology, primarily developed for civilian uses, into a matter of national security and sovereignty.[19] As a result, AI has become fully part of the contested global 'digital arms race,' raising major concerns about the broader risks associated with its use for offensive purposes.[20] This evolution is not surprising. It is in line with the broader securitisation of cyberspace over the past three decades, quickly, but not always correctly, associated with its militarisation in the discourse of states.[21]

With the rise of increasingly sophisticated and targeted state-sponsored cyberattacks since the late 2000s, cyberspace emerged as an imperative of securitisation and a new warfighting domain that required the mobilisation of exceptional means.[22] The representation of cyberspace as predominantly a threat to national security is not self-evident given the complex challenges in this domain, such as those posed by criminal organisations to individual interests that can equally hurt the security of end users and the security and stability of cyberspace itself.[23] Other characterisations that could have prevailed such as economic risk, criminal danger, or threats to individual user privacy have increasingly taken a back seat to national and international security concerns.[24] In the words of internet governance scholar, Milton Mueller 'cybersecurity is eating internet governance' and is pushing out alternative framings.[25] The security frame has progressively extended to all the digital technologies that could be weaponised in the context of digital warfare,

including AI, and drives international competition over digital technologies. This competition is both embodied and increasingly shaped by the fierce competition between the United States and China over the production, control, use and governance of digital technologies. Adam Segal argues that during the 1990s and 2000s the integration of the Chinese and American economies was perceived as mutually beneficial, both politically and economically, political decision-makers now consider that the risks outweigh the benefits.[26] And in both state discourses, the issue of security is at the heart of the rivalry. It should be noted that China has launched a massive plan to become the world leader in AI by 2030, with a 150-billion-dollar industry.[27] That is, the talent war is on.

The leadership of a few countries in AI capabilities also reveals uncomfortable strategic dependencies for many other countries. It has triggered a debate in the European Union about the risk associated with these dependencies and the need for strategic autonomy to ensure digital sovereignty.[28] But advancing AI technology appears to be a limited policy option to address these issues. In Chapter 4**,** Simona Soare questions the role of AI to advance European strategic autonomy in the field of security and defence. She argues that the adoption of AI is a 'distraction' as it introduces additional layers of complexity in the European defence while not contributing significantly to Europe's strategic autonomy. On the one hand, the integration of AI in the EU decision-making processes and the conduct of operations is challenging because of the EU's internal functioning in the field of defence. On the other hand, the lack of industrial capabilities and the strategic dependencies towards other powers are real and likely difficult to overcome. In Chapter 5 Arun Mohan Sukumar similarly demonstrates that relying on AI can introduce risks and strategic dependencies, as shown in the case of emerging powers. The chapter examines the role of AI in the development of public services, through examples of the health sector in Brazil, India and Singapore. It shows how, while states are urged to enhance data transparency and to develop digital services for their population, they become exposed to new risks that could set back progress in the digitalisation of states' mission-critical systems for years. That is, they face a trade-off between furthering digitalisation and accepting more security risks, an instance that speaks to the importance of thinking about the AI-cyber nexus not only in technical/operational terms but also considering broader strategic implications.

Armed forces worldwide have also recognised the strategic relevance of the AI-cyber nexus and have similarly engaged in a profound digital transformation of their operations.[29] On the one hand, this has considerably increased their reliance on digital technologies and data. On the other, it has created new risks and vulnerabilities. Soldiers evolve in a new digital environment that profoundly transforms the way they operate and creates new challenges that are sometimes hard to fully comprehend and govern. In this environment, AI offers promising new capabilities to improve the quality of intelligence, situational awareness, the conditions of training, the ability to operate

remotely, the precision and autonomy of weapon systems and, most importantly, the speed and scope of action. As result, the race for AI is thus also a race for military power and superiority and, again, raises strategic problems that are intimately related to operational ones. This representation resonates with a vision deeply ingrained in the US military culture that technology can provide military superiority. In Chapter 6, Jeppe Jacobsen and Tobias Liebetrau argue that this vision goes back a long time before AI and has dominated US military discourse since the Second Offset Strategy of the 1970s. That is, AI represents an operational innovation more than a strategic one.

While providing further evidence to a presumed return of great powers competition, the military superiority approach also feeds the fears inspired by the technology and is a driver for developing offense over defence, to maintain superiority over the enemy. But AI-enabled cyber capabilities might also convey the idea of control that is difficult if not illusory in cyberspace, given the highly dynamic nature of this environment.[30] And it does not take into consideration the vulnerabilities and associated risks that AI technology also brings about. Indeed, with the digital transformation of societies and armed forces, the attack surface keeps increasing. And while AI can considerably improve defence, the emphasis placed on offense could be a source of risk. Jeppe Jacobsen and Tobias Liebetrau demonstrate that the cyber arms race is not just a competition between great powers for AI-enabled cyber capabilities but also a specific arms race between offensive and defensive cyber capabilities, powered by AI. Given the lessons from discussions on how militaries balance offense and defence in cyberspace, they conclude that AI-enhanced cyber offensive capabilities are likely to dominate. And yet AI can backfire in many ways. As our societies grow increasingly dependent on digital technologies, the securitisation of AI technology could have important spill-over effects on the overall level of cyber (in)stability. The ongoing race for data and its exploitation for strategic advantages further blurs the lines between military and civilian operations, with inextricable consequences for the private sector and civil society, raising new legal and normative challenges.

Normative and legal considerations

Stemming directly from the above-mentioned technical/operational and strategic/geopolitical considerations is the necessity of regulating the adoption and use of AI technologies in cyberspace. The development of cyber capabilities, on the one hand, and AI and its possible applications in cyber conflicts on the other, have posed a dilemma to states and other actors: they are interested in these new technologies – notably to enhance their own operational capabilities and strategic posture – but they are at the same time concerned about the potential consequences of these developments for international peace and security. This dilemma lies at the core of the third final part of this volume, which deals with the normative and legal questions raised by AI applications in cyberspace. To understand these, this section also

introduces the international processes in which these normative and legal discussions are embedded and become deeply intertwined with states' strategic considerations.

On international cybersecurity, the United Nations General Assembly adopted its first resolution on "*Developments in the field of information and telecommunications in the context of international security*" in December 1998. Since 2004, the United Nations General Assembly has established six successive Groups of Governmental Experts (GGE) on this topic. The first and the fifth GGE failed to adopt a consensus report, reportedly because of disagreement in the discussions on specific branches of international law. The impossibility of the fifth GGE to adopt a consensus report in June 2017 led to disagreement on how to proceed. In 2018, this resulted in the adoption of two concurrent resolutions and the creation of two parallel processes, with largely the same mandate. In addition to the sixth GGE, an Open-Ended Working Group (OEWG) was established. In 2020, a new OEWG was established which will last until 2025 while there is as of, yet no new GGE planned.[31] Moreover, since 2020, some States are advocating for a new process on this topic, a Program of Action (PoA) for advancing responsible state behaviour in cyberspace,[32] which was welcomed in principle in November 2022 by the UN General Assembly. The second, third, fourth and sixth GGE as well as the first OEWG were successful in adopting consensus reports.[33] These reports notably affirmed that international law is applicable to cyberspace and listed specific rules and principles of international law deemed particularly relevant in this context. They also listed 11 norms of responsible behaviour in cyberspace. Taken together these reports constitute a framework of responsible State behaviour in cyberspace, encompassing international law and non-binding norms but also capacity building and confidence-building measures. Interestingly in the context of this book, the development of AI applications has never been mentioned in the GGE or OEWG reports, despite it being discussed in the 2019–2021 rounds of negotiation. While the issue did not make the cut of the 2021 consensus reports it does feature in the so-called Chair's summary of the OEWG process in its section dedicated to 'Threats': "Pursuit of increasing automation and autonomy in ICT operations was put forward as a specific concern, as were actions that could lead to the reduction or disruption of connectivity, unintended escalation or effects that negatively impact third parties."[34] Moreover, both in the context of the UN negotiations and outside of it, states and other actors have started to voice concerns about the role of automation and autonomy in cyber operations.[35]

The discussions on the international security dimensions of AI have been focusing on the development of LAWS. This matter was introduced in 2013 in the agenda of the Meetings of High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons (CCW). After a few informal meetings, these discussions took a similar path as the ones on international cybersecurity, with the establishment a GGE in 2016 which adopted 11 guiding principles on LAWS in 2019.[36] Through

these principles, the GGE affirmed the applicability of international law and in particular international humanitarian law as well as a series of ethical and non-binding principles. Surprisingly, Cybersecurity is only briefly mentioned in the sixth principle as one of the "appropriate non-physical safeguards [that] should be considered [w]hen developing or acquiring new weapons systems based on emerging technologies in the area of lethal autonomous weapons systems."[37] There is, however, no mention of autonomous cyber capabilities. Even though the link between cyber security and AI has been made in the context of the OECD[38] and in the UNESCO *Recommendation on the Ethics of Artificial Intelligence*,[39] both of these documents steer clear of national and international security. So until now, 'cyber' and 'AI' seem to be ships passing in the night in the UN's first committee.

This 'absence' is at the heart of the third part of this volume. To navigate this vacuum at the international level. Taddeo, McNeish, Blanchard, and Edgar discuss in Chapter 7 the efforts to define ethical frameworks to guide the use of AI in the defence domain at the domestic level – through the case of the United Kingdom – and propose a possible framework, articulated around five principles: justified and overridable uses; just and transparent systems and processes; human moral responsibility; meaningful human control; and, finally, reliable AI systems. At the core of these ethical considerations are the matter of technological autonomy and the need for some form of human control, involvement, or override: again, where does the human fit in 'the loop'? Going back to the international level, in Chapter 8 Louis Perez navigates the different discussion streams at the UN on Cyber on the one hand and LAWS on the other, before discussing how the current approach to LAWS could also be applied to autonomous cyber operations. Reflecting on the definition of LAWS, this chapter addresses the vital question of whether autonomous cyber capabilities could be considered LAWS and thus be concerned by the discussions on international law and ethics taking place in the framework of the CCW. In Chapter 9, Jack Kenny focuses on a specific principle of international law, the principle of non-intervention, that has been discussed extensively by the GGE and the OEWG. Building on these discussions, as well as on the existing scholarship on the application of this principle in cyberspace, this chapter looks at the specific challenges raised by automation for this principle with a specific focus on its coercion requirement. By going back to one of the operational dilemmas discussed earlier, the chapter elucidates this normative discussion through the analysis of different examples related to the interference in electoral processes using cyber means with a certain degree of autonomy.

The third and last part of the volume shows that the debates at the UN level have a while to go before they will be able to meaningfully address the intersection between AI technology and conflict in cyberspace. There are several reasons for that, which are related to the technical/operational and strategic/geopolitical perspectives outlined earlier. For one thing, most of the richer and top-tier (cyber) military states are often reluctant to forego

new military possibilities that may turn out to be game changers.[40] Countries like the United States, Israel and Russia, which are actively developing LAWS are dragging their feet in the GGE negotiations. History does not provide much evidence of weapons being banned before they are used. Also, politically the level of trust between some of the main negotiating parties is at a low point at this moment. The United States are increasingly in an adversarial competition with China – which is one of the main contenders for the 'AI crown' – and since the Russian invasion of Ukraine many states are actively trying to sanction and isolate the Russian Federation. These are not ideal circumstances to discuss restraint as a governance mechanism when it comes to new military and cyber technology. Lastly, there is a mandate mismatch between the two UN processes. The UN GGE on LAWS – as the name indicates – explicitly focuses on a specific technology (AI) in relation to *weapons*. The UN GGE on cybersecurity focuses on *state behaviour* as the focal point for its recommendations and usually aims to be as technology neutral as possible. If a bridge is to be built between these processes it will have to be built on sound reasoning on how technology impacts on, or changes, state behaviour in cyber conflict. Questions like whether 'state control' only exists when there is meaningful human control or also exist when in case of 'system control,' and whether automated and/or autonomous cyber-attacks are or can be (in)discriminate[41] are likely to be at the heart of that. In other words, only by understanding the relationship between AI and conflict in cyberspace as a comprehensive phenomenon, and embedded in broader geopolitical conflicts, can the international community truly move forward with meaningful regulation.

Notes

1 See, for example, Thomas Rid, *Cyber War Will Not Take Place* (London: Hurst and Company, 2013); Erik Gartzke, "The myth of cyberwar: Bringing war in cyberspace back down to earth," *International Security* 38, no. 2 (2013): 41–73; Robert Chesney and Max Smeets, eds., *Deter, Disrupt, or Deceive. Assessing Cyber Conflict as an Intelligence Contest* (Washington, DC: Georgetown University Press, 2023).
2 Lucas Kello, *The Virtual Weapon and International Order* (New Haven and London: Yale University Press, 2017).
3 Patryk Pawlak, Eneken Tikk, and Mika Kerttunen, "*Cyber Conflict Uncoded*," EUISS Brief no. 7, European Union Institute for Security Studies, 7 April 2020.
4 See Richard J. Harknett and Max Smeets, "Cyber campaigns and strategic outcomes," *Journal of Strategic Studies* 45, no. 4 (2022): 534–567; and Chesney and Smeets, *Deter, Disrupt, or Deceive*.
5 John Perry Barlow, "A Declaration of the Independence of Cyberspace" (8 February 1996).
6 David J. Betz and Tim Stevens, *Cyberspace and the State: Towards a Strategy for Cyber-power* (Abingdon: Routledge, 2011).
7 See, for example: Rebecca Slayton, "What is the cyber offense-defense balance? Conceptions, causes, and assessment," *International Security* 41, no. 3 (2017): 72–109.

8 H. Lin and J. Kerr, "On cyber-enabled information warfare and information operations," in *Oxford Handbook of Cybersecurity* (Oxford: Oxford University Press, 2021).
9 Ronald Deibert and Louis W. Pauly, "Mutual entanglement and complex sovereignty in cyberspace," in *Data Politics: Worlds, Subjects, Rights* (London: Routledge, 2019); Christian Ruhl et al., *Cyberspace and Geopolitics: Assessing Global Cybersecurity Norm Processes at a Crossroads* (Washington, DC: Carnegie Endowment for International Peace, 2020); Daniel Deudney, "Turbo change: Accelerating technological disruption, planetary geopolitics, and architectonic metaphors," *International Studies Review* 20, no. 2 (2018); Nicole Perlroth, *This Is How They Tell Me the World Ends: The Cyberweapons Arms Race* (Bloomsbury: Bloomsbury Publishing, 2021).
10 See, for example: Jon R. Lindsay, *Information Technology and Military Power* (Ithaca, NY: Cornell University Press, 2020); and Avi Goldfarb and Jon R. Lindsay, "Prediction and judgment: Why artificial intelligence increases the importance of humans in war," *International Security* 46, no. 3 (2022).
11 Michael C. Horowitz, "Artificial intelligence, international competition, and the balance of power," *Texas National Security Review* 1, no. 3 (May 2018).
12 See Monica Kaminska, Dennis Broeders, and Fabio Cristiano, "Limiting viral spread: Automated cyber operations and the principles of distinction and discrimination in the grey zone," in *13th International Conference on Cyber Conflict: 'Going Viral'* (Tallinn: CCDCOE, 2021).
13 Corien Prins et al, *iGovernment* (Amsterdam: Amsterdam University Press, 2011).
14 Tim Stevens, "Knowledge in the grey zone: AI and cybersecurity," *Digital War* 1, no. 1 (2020); Finn Brunton, *Spam: A shadow history of the Internet* (MIT Press, 2013).
15 This question is also relevant for debates on attribution. On this topic, see Joseph M. Brown and Tanisha M. Fazal, "#SorryNotSorry: Why states neither confirm nor deny responsibility for cyber operations," *European Journal of International Security* 6, no. 4 (2021); for a wider legal perspective, see: R. Liivoja, M. Naagel, and A. Väljataga, *Autonomous Cyber Capabilities Under International Law* (Tallinn, Estonia: CCDCOE, 2019).
16 Noran Shafik Fouad, "The non-anthropocentric informational agents: Codes, software, and the logic of emergence in cybersecurity," *Review of International Studies* 48, no. 4 (2022); Clare Stevens, "Assembling cybersecurity: The politics and materiality of technical malware reports and the case of Stuxnet," *Contemporary Security Policy* 41, no. 1 (2020); Myriam Dunn Cavelty and Andreas Wenger, "Cyber security meets security politics: Complex technology, fragmented politics, and networked science," *Contemporary Security Policy* 41, no. 1 (2020).
17 Benjamin M. Jensen, Christopher Whyte, and Scott Cuomo, "Algorithms at war: The promise, peril, and limits of artificial intelligence," *International Studies Review* 22, no. 3 (2020); Paul Scharre, *Army of None. Autonomous Weapons and the Future of War* (New York: W.W. Norton & Company, 2018); Michael C. Horowitz, "The ethics & morality of robotic warfare: Assessing the debate over autonomous weapons," *Daedalus* 145, no. 4 (2016).
18 Stevens, "Knowledge in the grey zone," 166.
19 Lene Hansen and Helen Nissenbaum, "Digital disaster, cyber security, and the Copenhagen School," *International Studies Quarterly* 53, no. 4 (2009).
20 Heather M. Roff, "The frame problem: The AI "arms race" isn't one," *Bulletin of the Atomic Scientists* 75, no. 3 (2019).
21 Sergei Boeke and Dennis Broeders, "The demilitarisation of cyber conflict," *Survival* 60 no. 6 (2018).
22 Max Smeets, *No Shortcuts: Why States Struggle to Develop a Military Cyber-Force* (London: Hurst Publishers, 2022).

23 Nazli Choucri and David D. Clark, *International Relations in the Cyber Age: The Co-Evolution Dilemma* (Cambridge, MA: MIT Press, 2019).
24 Ben Buchanan, *The Cybersecurity Dilemma: Hacking, Trust, and Fear Between Nations* (Oxford: Oxford University Press, 2016).
25 M. Mueller, "Is cybersecurity eating internet governance? Causes and consequences of alternative framings," *Digital Policy, Regulation and Governance* 19, no. 6 (2017).
26 Adam Segal, "Une guerre froide fluide: Les Etats-Unis, la Chine et la guerre technologique," *Hérodote* 2022, no. 184–185 (2022); see also, Kai Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order* (Boston, MA: Houghton Mifflin, 2018).
27 Graham Webster et al., "China's plan to 'lead' in AI: Purpose, prospects, and problems," *New America Foundation* (1 August 2017).
28 Dennis Broeders, ed., *Digital Sovereignty: From Narrative to Policy?* (EU Cyber Direct, 2022); T. Christakis, *'European Digital Sovereignty': Successfully Navigating Between the 'Brussels Effect' and Europe's Quest for Strategic Autonomy* (Multidisciplinary Institute on Artificial Intelligence/Grenoble Alpes Data Institute, 2020); Benjamin Farrand and Helena Carrapico, "Digital sovereignty and taking back control: From regulatory capitalism to regulatory mercantilism in EU cybersecurity," *European Security* 31, no. 3 (2022).
29 See, for example: Lindsay, *Information Technology*; James Johnson, "The AI-cyber nexus: Implications for military escalation, deterrence and strategic stability," *Journal of Cyber Policy* 4, no. 3 (2019); Joe Burton and Simona R. Soare, "Understanding the strategic implications of the weaponization of artificial intelligence," in *11th International Conference on Cyber Conflict (CyCon): Silent Battle*, ed. T. Minárik, S. Alatalu, S. Biondi, M. Signoretti, I. Tolga and G. Visky (Tallinn: CCDCOE, 2019).
30 Martin C. Libicki, "Cyberspace is not a warfighting domain," *Journal of Law and Policy for the Information Society* 8, no. 2 (2012).
31 For an overview, see: Dennis Broeders, "The (im)possibilities of addressing election interference and the public core of the internet in the UN GGE and OEWG: A mid-process assessment," *Journal of Cyber Policy* 6, no. 3 (2021): 277–279.
32 Aude Géry and François Delerue, "A new UN path to cyber stability," *Directions* (6 October 2020); Valentin Weber, "How to strengthen the program of action for advancing responsible state behavior in cyberspace," *Just Security* (10 February 2022).
33 UNGA, *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*, UN Doc A/65/201 (30 July 2010); UNGA, *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*, UN Doc A/68/98 (24 June 2013); UNGA, *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*, UN Doc A/70/174 (22 July 2015); UNGA, *Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security*, UN Doc A/76/135 (14 July 2021); UNGA, *Report of the Open-ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security*, UN Doc A/75/816 (18 March 2021).
34 Chair of the OEWG, *Chair's Summary of the OEWG on Developments in the Field of Information and Telecommunications in the Context of International Security*, UN Doc AC.290/2021/CRP.3 (10 March 2021).
35 See Kaminska, Broeders, and Cristiano, "*Limiting Viral Spread*," 62–64.
36 CCW, *Guiding Principles affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System*, annexed (Annex III)

to the *Final Report* of the meeting of the high contracting parties to the convention on prohibitions or restrictions on the use of certain conventional weapons which may be deemed to be excessively injurious or to have indiscriminate effects, 19 December 2019, CCW/MSP/2019/9, Annex III, 10.

37 Ibid., principle (f).

38 OECD, *Recommendation of the Council on Artificial Intelligence* (Paris: OECD, 2019).

39 UNESCO, *Recommendation on the Ethics of Artificial Intelligence* (UNESCO, 2021).

40 See, for example: John Arquilla, *Bitskrieg. The New Challenge of Cyberwarfare* (London: Polity, 2021).

41 Kaminska, Broeders and Cristiano, "*Limiting Viral Spread*."

Bibliography

Arquilla, John. *Bitskrieg. The New Challenge of Cyberwarfare*. London: Polity, 2021.

Barlow, John Perry. *A Declaration of the Independence of Cyberspace*. 8 February 1996. http://www.eff.org/~barlow/Declaration-Final.html.

Betz, David J., and Tim Stevens *Cyberspace and the State: Towards a Strategy for Cyber-Power*. Abingdon: Routledge, 2011.

Boeke, Sergei, and Dennis Broeders. "The demilitarisation of cyber conflict." *Survival* 60, no. 6 (2018): 73–90. https://doi.org/10.1080/00396338.2018.1542804.

Broeders, Dennis. "The (im)possibilities of addressing election interference and the public core of the internet in the UN GGE and OEWG: A mid-process assessment." *Journal of Cyber Policy* 6, no. 3 (2021): 277–297. https://doi.org/10.1080/23738871.2021.1916976.

Broeders, Dennis, ed. *Digital Sovereignty: From Narrative to Policy?* EU Cyber Direct, 2022.

Brown, Joseph M., and Tanisha M. Fazal. "# SorryNotSorry: Why states neither confirm nor deny responsibility for cyber operations." *European Journal of International Security* 6, no. 4 (2021): 401–417.

Brunton, Finn. *Spam: A Shadow History of the Internet*. MIT Press, 2013.

Buchanan, Ben. *The Cybersecurity Dilemma: Hacking, Trust, and Fear Between Nations*. Oxford: Oxford University Press, 2016.

Burton, Joe, and Simona R. Soare. "Understanding the strategic implications of the weaponization of artificial intelligence." In *2019 11th International Conference on Cyber Conflict (CyCon): Silent Battle*, edited by T. Minárik, S. Alatalu, S. Biondi, M. Signoretti, I. Tolga and G. Visky, 1–17. Tallinn: CCDCOE, 2019. https://doi.org/10.23919/CYCON.2019.8756866.

CCW. *Guiding Principles Affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System*. CCW/MSP/2019/9, 19 December 2019.

Chair of the OEWG. *Chair's Summary of the OEWG on Developments in the Field of Information and Telecommunications in the Context of International Security*. UN Doc AC.290/2021/CRP.3, 10 March 2021.

Chesney, Robert, and Max Smeets, eds. *Deter, Disrupt, or Deceive. Assessing Cyber Conflict as an Intelligence Contest*. Washington, DC: Georgetown University Press, 2023.

Choucri, Nazli, and David D. Clark. *International Relations in the Cyber Age: The Co-Evolution Dilemma*. Cambridge, MA: MIT Press, 2019.

Christakis, T. *'European Digital Sovereignty': Successfully Navigating between the 'Brussels Effect' and Europe's Quest for Strategic Autonomy*. Multidisciplinary Institute on Artificial Intelligence/Grenoble Alpes Data Institute, 2020.

Deibert, Ronald, and Louis W. Pauly. "Mutual entanglement and complex sovereignty in cyberspace." In *Data Politics: Worlds, Subjects, Rights*, edited by Didier Bigo, Engin Isin and Evelyn Ruppert, 81–88. London: Routledge, 2019.

Deudney, Daniel. "Turbo change: Accelerating technological disruption, planetary geopolitics, and architectonic metaphors." *International Studies Review* 20, no. 2 (2018): 223–231.

Dunn Cavelty, Myriam, and Andreas Wenger. "Cyber security meets security politics: Complex technology, fragmented politics, and networked science." *Contemporary Security Policy* 41, no. 1 (2020): 5–32.

Farrand, Benjamin, and Helena Carrapico. "Digital sovereignty and taking back control: From regulatory capitalism to regulatory mercantilism in EU cybersecurity." *European Security* 31, no. 3 (2022): 435–453. https://doi.org/10.1080/09662839.2022.2102896.

Fouad, Noran Shafik. "The non-anthropocentric informational agents: Codes, software, and the logic of emergence in cybersecurity." *Review of International Studies* 48, no. 4 (2022): 766–785.

Gartzke, Erik. "The myth of cyberwar: Bringing war in cyberspace back down to earth." *International Security* 38, no. 2 (2013): 41–73.

Géry, Aude, and François Delerue. "A new UN path to cyber stability." *Directions*(6 October 2020). https://directionsblog.eu/a-new-un-path-to-cyber-stability/.

Goldfarb, Avi, and Jon R. Lindsay. "Prediction and judgment: Why artificial intelligence increases the importance of humans in war." *International Security* 46, no. 3 (2022): 7–50.

Hansen, Lene, and Helen Nissenbaum. "Digital disaster, cyber security, and the Copenhagen school." *International Studies Quarterly* 53, no. 4 (2009): 1155–1175.

Harknett, Richard J., and Max Smeets. "Cyber campaigns and strategic outcomes." *Journal of Strategic Studies* 45, no. 4 (2022): 534–567.

Horowitz, Michael C. "The ethics & morality of robotic warfare: Assessing the debate over autonomous weapons." *Daedalus* 145, no. 4 (2016): 25–36. https://doi.org/10.1162/DAED_a_00409.

Horowitz, Michael C. "Artificial intelligence, international competition, and the balance of power." *Texas National Security Review* 1, no. 3 (May 2018): 36–57.

Jensen, Benjamin M., Christopher Whyte, and Scott Cuomo. "Algorithms at war: the promise, peril, and limits of artificial intelligence." *International Studies Review* 22, no. 3 (2020): 526–550.

Johnson, James. "The AI-cyber nexus: Implications for military escalation, deterrence and strategic stability." *Journal of Cyber Policy* 4, no. 3 (2019): 442–460. https://doi.org/10.1080/23738871.2019.1701693.

Kaminska, Monica, Dennis Broeders, and Fabio Cristiano. "Limiting viral spread: Automated cyber operations and the principles of distinction and discrimination in the grey zone." In *13th International Conference on Cyber Conflict:* '*Going Viral*,' edited by T. Jančárková, L. Lindström, G. Visky and P. Zotz, 59–72. Tallinn: CCDCOE, 2021.

Kello, Lucas. *The Virtual Weapon and International Order.* New Haven, CT and London: Yale University Press, 2017.

Lee, Kai Fu. *AI Superpowers: China, Silicon Valley, and the New World Order.* Boston, MA: Houghton Mifflin, 2018.

Libicki, Martin C. "Cyberspace is not a warfighting domain." *Journal of Law and Policy for the Information Society* 8, no. 2 (2012): 321–336.

Liivoja, R., M. Naagel, and A. Väljataga. *Autonomous Cyber Capabilities under International Law*. Tallinn, Estonia: CCDCOE, 2019.

Lin, H., and J. Kerr. "On cyber-enabled information warfare and information operations." In *Oxford Handbook of Cybersecurity*, edited by P. Cornish, 251–272. Oxford: Oxford University Press, 2021.

Lindsay, Jon R. *Information Technology and Military Power*. Ithaca, NY: Cornell University Press, 2020.

Mueller, M. "Is cybersecurity eating internet governance? Causes and consequences of alternative framings." *Digital Policy, Regulation and Governance* 19, no. 6 (2017): 415–428. https://doi.org/10.1108/DPRG-05-2017-0025.

OECD. *Recommendation of the Council on Artificial Intelligence*. Paris: OECD, 2019.

Pawlak, Patryk, Eneken Tikk, and Mika Kerttunen. *Cyber Conflict Uncoded*. EUISS Brief no. 7, European Union Institute for Security Studies, 7 April 2020. https://www.iss.europa.eu/content/cyber-conflict-uncoded.

Perlroth, Nicole. *This is How They Tell Me the World Ends: The Cyberweapons Arms Race*. Bloomsbury: Bloomsbury Publishing, 2021.

Prins, Corien, Dennis Broeders, Henk Griffioen, Anne-Greet Keizer, and Esther Keymolen. *iGovernment*. Amsterdam: Amsterdam University Press, 2011.

Rid, Thomas. *Cyber War Will Not Take Place*. London: Hurst and Company, 2013.

Roff, Heather M. "The frame problem: The AI "arms race" isn't one." *Bulletin of the Atomic Scientists* 75, no. 3 (2019): 95–98.

Ruhl, Christian, Duncan Hollis, Wyatt Hoffman, and Tim Maurer. *Cyberspace and Geopolitics: Assessing Global Cybersecurity Norm Processes at a Crossroads*. Washington, DC: Carnegie Endowment for International Peace, 2020.

Scharre, Paul. *Army of None. Autonomous Weapons and the Future of War*. New York: W.W. Norton & Company, 2018.

Segal, Adam. "Une guerre froide fluide: Les Etats-Unis, la Chine et la guerre technologique." *Hérodote* 2022, no. 184–185 (2022): 271–284.

Slayton, Rebecca. "What is the cyber offense-defense balance? Conceptions, causes, and assessment." *International Security* 41, no. 3 (2017): 72–109. https://doi.org/10.1162/ISEC_a_00267.

Smeets, Max. *No Shortcuts: Why States Struggle to Develop a Military Cyber-Force*. London: Hurst Publishers, 2022.

Stevens, Clare. "Assembling cybersecurity: The politics and materiality of technical malware reports and the case of Stuxnet." *Contemporary Security Policy* 41, no. 1 (2020): 129–152.

Stevens, Tim. "Knowledge in the grey zone: AI and cybersecurity." *Digital War* 1, no. 1 (2020): 164–170.

UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. UNESCO, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000379920.

UNGA. *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*. UN Doc A/65/201, 30 July 2010.

UNGA. *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*. UN Doc A/68/98, 24 June 2013.

UNGA. *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*. UN Doc A/70/174, 22 July 2015.

UNGA. *Report of the Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security*. UN Doc A/75/816, 18 March 2021.

UNGA. *Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security*. UN Doc A/76/135, 14 July 2021.

Weber, Valentin. "How to strengthen the program of action for advancing responsible state behavior in cyberspace." *Just Security*, 10 February 2022. https://www.justsecurity.org/80137/how-to-strengthen-the-programme-of-action-for-advancing-responsible-state-behavior-in-cyberspace/.

Webster, Graham, Rogier Creemers, Paul Triolo, and Elsa Kania. "China's plan to 'lead' in AI: Purpose, prospects, and problems." *New America Foundation*, 1 August 2017. https://www.newamerica.org/cybersecurity-initiative/blog/chinas-plan-lead-ai-purpose-prospects-and-problems/.

Part I

Technical and operational challenges

2 The unknowable conflict

Tracing AI, recognition, and the death of the (human) loop

Andrew C. Dwyer

Situating unknowability

From drones, cybersecurity, to robotics and beyond, international conflict is intricately intertwined with computation, articulating norms that increasingly consist of calculation rather than social negotiation.[1] As computation arguably unsettles the dominance of social norms and decision across our societies, from social welfare, policing, border points to social media, these social negotiations are ever-more intensively (re)asserted as central in attempts to limit undesirable[2] outputs from 'artificial intelligence' (AI) systems.[3] The 'human in the loop' thus steps in to assert authority and control as an exterior, and rational, residue of human-centric decision that can be drawn upon to restrain AI in international conflict and elsewhere.[4] This chapter argues that a dependency by governments, militaries, and corporations on socio-normative decision through the human in the loop must be questioned, when such socio-normativity is being transformed, modified, and reconstituted by the recursive calculability of computation. By this, I mean the ability of computation to iteratively reassess and identify new, calculative, connections and relations. This has implications not only for the *outputs* of contemporary AI systems but also for how international conflict unfolds and its ethico-political accounts of practice, control, and authority.

Voracious debate has ensued across academia and policy, from AI Ethics[5] to autonomous weapons systems.[6] Almost all experiment with the human as part of a conditioning loop of control and ethical restraint for AI systems. In more nuanced readings, humans are situated at multiple points of decision to permit intervention at crucial points in time – as events – where norms, responsibility, and ethics can be located and rendered accountable. However, as I shall argue, as machine learning algorithms and other forms of 'automated' and 'autonomous' technologies become commonplace, there is a receding possibility for appropriate translations, or 'explainability.'[7] This is because AI systems work on an alternative form of calculative norm-making that is not aligned to human commensurability through their architectural and logical modes of (re)cognition. My claim is that, due to this change in how norms are formed through and with computation, international conflict

DOI: 10.4324/9781003284093-3

progressively becomes quintessentially unknowable. It is unknowable *where* conflict takes place, *how* it is recognised by computation, and *who* should take certain actions.

In exploring AI systems and unknowability in contemporary, and probable future, conflict, I probe 'deep' reinforcement learning (RL). Deep RL can be summarised, for now, as an iterative system that includes 'agents' who 'learn' from an environment to improve their next action to attain a (pre-determined and desired) goal. I do this as a speculative exercise due to its strong applicability to adversarial moves by 'agents' to examine their potential application in both military wargaming as well as in offensive cyber operations. In doing so, I use these cases to raise questions over the neglected role of AI on international conflict beyond physical violence and lethality. Their application to conflict, in contrast to robotics for instance, may be understood as 'distanced' from the frontline in the case of military wargaming or unlikely to lead to a lethal outcome[8] in the case of offensive cyber operations. I will instead argue that, due to the way that norms are produced through the cognitive practices of reading, interpretation and action by computation and AI systems, imperceptible changes occur not only in how these become practiced but also how they are themselves formed and become *optimised* within the contours of international conflict.

Unknowability is, of course, by no means new; there have always been obscurities, where technologies have offered the promise to illuminate the spaces of conflict, only this time through big data and algorithmic processing. Computation, I shall contend however, is not simply another technology,[9] but actively takes part in forming what the threat, the adversary, and indeed what norms come to be established. AI systems are full of doubt, as much as Clausewitz noted this with regards to war,[10] in how they readily provide a 'solution' that embeds many assumptions and connections unknown through both their construction and their 'learning data.'[11] Drawing on Michel Foucault's inversion of Clausewitz "that politics is the continuation of war by other means,"[12] AI systems offer new ethico-political formations that condition discussions on 'grey zone' warfare,[13] as well as challenging notions of where conflict is conducted and what the international means as it disrupts and bends its spatial attributes across our societies. This chapter then outlines the 'unknowable conflict' that is made present through, and by, AI systems through an exploration of deep RL.

By questioning a reliance of decision – not as a claim that humans have *no responsibility*, but rather shared and complicated by computational (re)cognition and choice, I thus: (1) outline how technology and conflict has been predicated on the notion of the tool which can be controlled and rendered knowable; (2) detail how computational capacities for (re)cognition permit certain choices that are incommensurable to social negotiation and norms; (3) delve into 'deep' RL using military wargaming and offensive cyber operations as exemplars to consider how they may transform conflict; (4) this then informs a discussion on the ethico-political implications of tracing conflict

when computational recognition diverges from our knowability, and the consequences of conflict amid the loss of the face of the other; before (5) concluding that the human in the loop is unsuitable for ethico-political accounts of AI systems, where computation is transforming norms and practices, and in the process decomposing the notions of international conflict itself.

Technologies and control

Conflict has always been dependent on various non-digital technologies, only more recently have computation and AI systems introduced recursive capacities. These challenge conventional notions of an authoritative human and a subservient technological tool. This has worked on embedded assumptions of technology as lacking in complexity, as inherently deconstructable, and therefore knowable, compared to the sophisticated impenetrability of human thought. Yet, since at least the 1950s, an equivalence has been made between computation and human cognition and intelligence (such as through the neuroscientific inspiration for contemporary machine learning architectures[14]). This has led to an uneasy comparison between *our* intelligence as able to be usurped by 'machines' but one that cannot be socially nor politically equivalent. This has animated cybernetic debates over automation and warfare.[15] Yet, AI systems can be considered neither 'artificial' nor 'intelligent' but rather something distinct and alternate, as I shall explore in greater depth in the next section. AI systems have disturbed dominant perspectives between human control, agency, and intent over computation as *tool* as they have grown in recursive capacity.[16] AI systems may perform in 'unexpected' ways, without direct human oversight, or not be easily rendered knowable to decision-makers and their norms, but still be based on a wholly logical outcome of their prepositions.[17] In this chapter, I do not dwell on the divide between automation and autonomy,[18] but rather emphasise computational forms of recognition and its impact.

Control, likewise, has been a persistent theme in relation to technology and conflict, which has emerged most substantially in discussions on the potential 'autonomy' of lethal decisions of autonomous weapons systems (LAWS).[19] Most academic and popular writing argues that lethal capabilities should not be performed by computation without explicit human authorisation. To address this within militaries, at least within the Anglosphere, 'human-machine teaming'[20] and 'centaurs'[21] have been proposed to mix the processing capabilities of AI with human oversight, that keep the human in or on the loop to varying degrees. Such a position has been criticised over the possibility to have 'meaningful' control over LAWS, when there may be not enough information or time to consider a response.[22] Such concerns over lethal capabilities were most recently brought to attention in a UN report on the use of drones in Libya.[23] However, control based on *the loop* of the human as in, on, or even out, of some form of recursive and neat diagram is based on a certain understanding to technology and the role of human control in

such situations. That is, a human that can *rationally execute* their decision and judgement, either through the code that they write, the operators who *control* the technology, or strategic decision-makers who assess the benefits of certain engagements.

This position assumes a technological tool that is knowable, yet with AI systems, the construction, architecture, and learning data bring together new formations that challenge such a knowability of the context. Technological assessments of control and authority often present a one way or cyclical flow of intent from human to AI system, where there is an "outside to the algorithm – an accountable human subject who is the locus of responsibility, the source of a code of conduct with which algorithms must comply."[24] Such perspectives, as political geographer Louise Amoore argues, stem from a knowing, responsible, accountable human. Whereas prior technologies, such as non-computational nuclear weapons, may exhibit roughly deterministic outcomes, many AI systems are recursive and non-linear, and thus demand different attention.[25] As practitioner Keith Dear says, contemporary computation has the capacity for a "deeper recursive reasoning that humans are capable of, spotting patterns that humans cannot see, and which are beyond narrative description, without the encumbrance of human emotional and cognitive limitations."[26] Likewise, media theorists have argued for some time that there is no direct linear connection between highly recursive systems and the 'performance' of software.[27] There thus appears to be a growing convergence of diverse disciplinary and practitioner perspectives over the difficulty of control, and the performative qualities, of computation.

Although AI does not represent "some kind of epistemological break with past modalities … it reanimates and extends a longstanding intertwining of the practices and techniques of security with computational processes,"[28] however, recursion does represent one important change. Technological rationalities of conflict based on control, knowability, and authority have been gradually eroded by computation, and more recently through the recursive cognitive capacities of AI systems. Yet, anthropomorphic narratives of AI remain, with words such as "*smart*, *intelligent*, *sentient*, and *thinking*"[29] permitting a particular resonance to human cognitivist comparison. Computation, in this reading, operates in a similar way to our forms of intelligence. This collapses different forms of recognition between computation and social negotiation, so that the former can be understood as simply an automation (or even the mind as a computer).[30] This leaves AI systems sitting in an awkward position between being able, because they are calculative and logical, to be controlled, and yet precisely because of their recursion, they are compared to the human mind.

Recognising conflict

As I have discussed, AI systems have disrupted notions of control, primarily through the role of recursivity. Such a recursivity is built on an increasing

abstraction from digital binary upon which computation is built.[31] As N. Katherine Hayles writes in *Unthought*,[32] we are increasingly mediated through 'cognitive assemblages' which imbricate us with computation in ways that may not be immediately evident. This is because computation has a capacity for choice-making through its cognitive capacities that permits it to *read*, *interpret*, and crucially *act*, upon signs. This can be seen in the basic 'reading' of code across different programming languages, interpreting this calculatively and through logical axioms, and then taking an action based on this interpretation. This can appear very limited at the 'lower' levels of computation, such as at binary logic gates, and in tightly defined programming, yet once recursivity, approximation, and probability come into the picture, the choices available to computation through its recursive recognition grow. In a similar vein, we can understand that this recursivity has permitted a move from deductive to inductive logic that AI systems work upon, which Luciana Parisi summarises as *abduction*, which is a "speculative dimension of reasoning."[33] Thus, this chapter engages in a similar speculative exercise to explore the implications of deep RL for international conflict.

Computational choice-making is thus not made on the same basis as humans. It is an alternative form of abstraction through digital binary that is reconstituted into higher-level recursive connections and relations. Philosopher Beatrice Fazi argues that this leads to a quintessential incommensurability between these two forms of recognition.[34] This is because humans cannot adequately recognise and translate the sheer complexity of the high dimensional relations in deep RL, with thousands, if not more, potential options. Computational choices, of course, are not equivalent to the highly reflexive decisions that humans make. Yet, deep RL cannot abstract and translate the immense possibilities of all the different options, leading to outputs that are often singular but arrived at from numerous data points that Amoore articulates as an algorithmic aperture, which is "an opening that is simultaneously a narrowing, a closure."[35] Although computational technologies have been integrated into conflict for several decades, 'automation' has posed fewer concerns as it did not appear to substantively diverge from human intent and control. It is only with greater abstraction and recursivity that has brought it to greater attention, but choice has always been a condition of computation. It is only with AI systems today that this cannot be discarded as 'error' or 'environmental complexity.'

To deal with incommensurability of computational recognition and choice, governments, militaries, and corporations have therefore sought to locate the human in *the loop* as a resolution to the problem. In deep RL, one could argue that the process of setting of a goal for an artificial agent means that they adhere to 'social norms.' However, within deep RL, there is significant interpretation by computation to approximate values and optimise towards a goal. Although the methods may appear to be zoning in on social norms, computation does not recognise that an action may be considered unethical, be good, bad, or any other normative standard that may be understood by humans. Computation's

choices do not explicitly align with social norms, although they often can, producing outputs that are not intuitive to those who are creating strategy or launching an offensive cyber operation. Conflict in this sense could then become reliant, through computational recognition and choice, on approximation, probabilities, and inductive methods of sensing the world. This is not to say that *the loop* could not come in and ask critical questions over such a position, but it is about the more insidious and subtle transformations that may occur. It could slowly transform what is considered as competent behaviour in military organisations, or indeed, as more 'effective' and 'cost-efficient.'

Indeed, similar methods are being pursued by militaries – such as through agent-based modelling in the UK – to help shape strategic decision-making. It does not take a leap of imagination that in the future, if deep RL is substantively developed, vulnerability scanning, new investments, and new strategy and operations will be carried out through computational optimisation. Yet, even if a 'meaningful' human is present, computation is still *making choices* on big data in deep RL through an approximation of rewards in an environment. This raises serious questions over the power of such choices and how they intersect with decisions made by computation in the future. As Suchman, Follis, and Weber argue, this could further contribute to a predictive technoscience.[36] This could have damaging effects on potential escalation as militaries intersect with multiple other AI systems. Thus, RL may expand conflict as it simultaneously retreats to cloud servers away from conventional spaces of conflict. As Mackenzie and other social theorists have contended,[37] there is not a simple way with which humans recognise the world, process it, and then apply this to computation. Instead, there is a complex mediation between multiple humans and computers – cognitive assemblages – which mean it is not easy to say how conflict will morph in the future as greater capacities for recognition are provided to computation and AI systems.

The loop (human-machine teaming and centaurs), as I explored previously, can be considered as *posthuman* subjects that are grafted on to a liberal notion of humans as atomistic and rational beings.[38] Instead, there must be an emphasis on shared notions of action, thinking, and decision that arrive with computation and more extensively in architectures of international conflict. As Elke Schwarz notes in discussions on the role of LAWS, there must be a consideration of

> autonomous weapons not as discrete devices programmed to produce a pre-determined outcome, but as cognitive assemblages, a complex of sensors, information networks, transmitters, and hardware, which offer affordances, together with various human designers, and operators, that shape our very ideas of what it means to exert moral agency.[39]

This suggests that recognising conflict is not simply one that *belongs* to us alone. Rather, conflict is shared and built together in ways that rely on AI systems to articulate and render options anew.

As Tim Stevens reflects,

> [t]he disconnect between humans and the speed of networked computing machines means that the absolute speeds of communication can never truly be known to the unaided observer and leads to ever-greater reliance on computers as the providers of security.[40]

This speed afforded by AI is now also complicated by greater recursive capacities afforded by computational recognition and its choices. This recognition and choice-making by computation is however not working in discrete ways, as in one algorithm to a human operator. AI systems are interacting with one another, creating whole cognitive ecologies that further increase the incommensurability of the future of conflict. This must be processed within bureaucracies, organisations, and how they develop AI systems and integrate these in military platforms.[41] This may permit greater insight, greater forms of recognition, but not ones that may be sensible to our knowledge, practices, and ways of living. Greater recursive recognition, albeit perhaps unintuitively, is generating a lesser form of awareness; the fog of war grows denser.

Deep reinforcement learning

Before considering at greater depth the implications of computational forms of recognition upon decisions, norms, and the ethico-political status of conflict, I here develop an empirical appreciation of 'deep' RL across military wargaming and offensive cyber operations. RL is considered as one branch of machine learning in addition to two others – supervised and unsupervised learning. In supervised learning 'features' are labelled and defined by the designer, permitting greater (human and social) categorisation, whereas in unsupervised learning, such features are chosen by the algorithm without explicit human intervention (although with significant influence on the weights of the model). In the simplest of terms, if there is a set of photographs of cats, supervised learning will be given features with labelled data to make connections, whereas in unsupervised learning, the algorithm will create its own features to identify cats without labelled data.[42] These different forms – supervised, unsupervised, and RL – have come to be collectively understood as machine learning (algorithms).

Machine learning has become dominant in recent advances in computer science due to its significant capacities to recognise the world through recursive methods as I have substantively developed. As leading computer scientist Yann Le Cun writes in *Quand la machine apprend* [When the machine learns], the emergence of artificial neural network algorithms, in particular, have been key to permitting complex tasks of recognition to be conducted, ranging from object recognition, to enabling autonomous vehicles, through to natural language processing.[43] Such methods have permitted a complex threading of 'big' data that appear to offer the capacity for reasoning and

decision-making. As philosopher Yuk Hui details, "[c]ontrary to automation considered as a form of repetition, recursion is an automation that is considered to be a genesis of the algorithm's capacity for self-positing and self-realization."[44] Machine learning algorithms then appear to replicate some of the useful qualities of humans, recursively analysing data, and in many cases do this with a capacity that far exceeds human capabilities.

RL algorithms meanwhile came to popular attention when Google's DeepMind revealed *AlphaGo Zero* that "is no longer constrained by the limits of human knowledge."[45] This could beat expert players of the game Go by offering a *generalised* method for beating players of many other games.[46] This was dependent on a then-new method, 'Deep Q-Networks' (DQNs), that combine RL with unsupervised machine learning – in this case convolutional neural networks.[47] RL can be most simply expressed as composed of an agent that *optimises* its actions to gain a reward within an environment through trial and error. Crucial is the generation of an *appropriate* value for the reward to permit 'new' experimental actions, as well as maintaining a record and reward signal of what has worked in the past. The aim of RL is to improve a policy so that it can maximise the reward for the agent, and iteratively and recursively improve its strategy. As one computer science text notes, "an agent must be able to learn from its own experience."[48] That means that an agent can recursively iterate what may be the optimal strategy to achieve a defined goal.

However, in RL, there are limits to the number of different potential environmental 'states' that can be processed, which leads to the 'curse of dimensionality.'[49] This is where there are too many states to be processed, with higher dimensionalities (or simply connections) between data, which make it computationally infeasible in both processing power and time. Another difficulty is providing a reward that both appreciates short-term gain alongside long-term exploration of different potential actions. When 'large' problems are addressed (as in the case of conflict, with its many environments and adversary actions), it becomes infeasible to rely on hard-coded reward scenarios for environments that provide a very high number of potential options. Thus, DeepMind's DQNs offer a capacity to approximate a reward using 'experience replay' to allow an agent to experiment and optimise their actions. Deep RL "uses neural networks to represent the policy and/or the value function, which can then approximate arbitrary functions and process complex inputs."[50] For conflict scenarios, the use of deep RL offers the promise to bridge "the divide between high-dimensional sensory inputs and actions."[51] Yet, as deep RL grows in its recursive and abstractive capacities through calculative cognition, the more it must rely on approximation and probability.

There are many different flavours of deep RL. This can include those using a predictive model and those that are model-free (where the learning algorithm develops a control policy directly) or have different methods to develop 'self-play' to train their agents and policies. Indeed, supervised and unsupervised learning "do not provide a principle for action selection and therefore cannot in isolation be enough for goal-orientated intelligence"[52]

unlike RL. This emphasis on goal orientation is purported by those who work in the area as key to more 'general' AI and the benefits for goal-oriented machine learning are clearly of interest to militaries. The capacity for action selection, along with possibilities for multi-agent RL, has led to its application in autonomous air traffic control[53] and predator avoidance in robotic systems.[54] There are also connections between game theory and deep RL[55] that make it applicable to a wide range of applications, including military wargaming and offensive cyber operations.

Wargaming for conflict

Wargaming has a long and complex history,[56] has been widely applied to cyber conflict,[57] as well as in assessments of how AI systems, including deep RL, can inform such practices.[58] Due to the capacity of deep RL to offer many environmental states and goal-driven agents, they appear to be particularly suitable to wargaming. Unlike a focus on 'human-machine teaming' and other hybrid forms of human decision making in conflict in the UK and USA, China anticipates that "AI will accelerate the process of military transformation, causing fundamental changes to military units' programming, operational styles, equipment systems and models of combat power generation."[59] The drive to work with AI systems in military settings, such as thinking around the role of 'improving' planning and strategy, could use deep RL to test different scenarios and adversary actions.

Deep RL could deeply inform how conflict is simulated and the forms of normative, and calculative, knowledge that are produced. Even today, 'agent-based' modelling[60] is central to UK Strategic Command's *Digital Strategy for Defence*.[61] This includes continued funding for a Single Synthetic Environment procured from the UK business, *Improbable*,[62] with a claimed ability for the UK to model responses to potential conflict scenarios. Although it is unclear from public statements whether (deep) RL plays a part, it is likely to incorporate at least some parts of an AI system. Likewise, researchers from Tianjin University in China have been experimenting with deep RL to aid in wargaming,[63] demonstrating that these forms of AI systems are already being tested for their application to such activities.

In a report commissioned for the UK Defence Science and Technology Laboratory (Dtsl), there is an extensive reflection on the capacity for the integration of deep RL into wargaming.[64] The report claims that "there is little work showing good performance with Deep RL on unseen environments … [t]his does not mean Deep RL is not useful in the context of wargames, but is not suitable for completely unseen scenarios."[65] This is affirmed by computer scientist Le Cun, who notes that deep RL

> simulators have to be powerful and precise enough, that is to say, they must reflect what is happening in reality enough so that, once the system is simulated, its capability can be transferred to the real world. This is not always possible.[66]

Although today deep RL may not be applied well in 'unseen environments,' this may be becoming less of a limiting factor. In 2019, DeepMind developed a multi-agent deep RL, *AlphaStar*, to play at human 'grandmaster' levels in the complex strategy game *StarCraft II*.[67] Due to the limited observations from the viewpoint of the agent in *StarCraft II*, as well as unseen environments and adversary movements, this is distinct to the bounded state solutions of *AlphaGo*.[68] The authors argue that "many applications have complex strategy spaces that contain cycles or hard exploration landscapes, and agents may encounter unexpected strategies or complex edge cases when deployed in the real world."[69] For wargaming, this ability to work with a partiality of information and obscured environments is important. Yet, as Goodman, Risi, and Lucas note, various deep RL are unlikely to be used in planning for operations but they could be useful for concept and force developments due to their "explicitly exploratory goals."[70]

Although it is unlikely that deep RL will be used for conflict operations at least in the near future, its exploratory capacity could be adopted for strategy in ways that are more advanced, yet recursive, than agent-based modelling. The capacity to model adversary behaviour could shift how militaries understand investment decisions, strategic planning, as well as the potential future possibilities of adversary behaviour. Although individuals may dispute certain outcomes of deep RL, military bureaucracies and other agencies may be tempted to subtly rely on such outputs. It is well documented that military personnel typically trust the outputs of computation, at the very least in time critical situations.[71] However, it is my argument that such recursive techniques of deep RL could transform what is considered normal in conflict, and actively shape how environments and procurements are made. It is not about individual, rational decision makers, but how such practices of deep RL, may shape organisational cultures, slowly revolutionising the practice of strategy. Changes may be imperceptible on a case-by-case basis but relying on deep RL's approximation of behaviour leads to an *optimisation* of conflict itself, with ethico-political consequences that could embolden strategic actions or prioritise subtle manoeuvres, each with potentially worrying complications.

Offensive cyber operations

Unlike in wargaming, the application of deep RL may be more difficult in operational 'kinetic' settings – such as with drones and robotics – due to the large range of unseen and data-poor environments.[72] Offensive cyber operations, by contrast, may be more readily adopted due to their lack of connection with lethality and therefore making it a space of 'innovation.' In *Automating Cyber Operations*, Buchanan et al. note how RL "could someday be a game-changer for offensive cyber operations,"[73] despite much contemporary work focusing on its application to defensive methods, such as protecting the 'Internet of Things' and improving intrusion detection systems.[74]

One development however includes Baillie et al.'s prototype autonomous cyber operations 'gym' that tests operations across 'blue' and 'red' adversarial teams.[75] They claim that it "aims to provide a scalable, efficient and flexible training environment that uses a high fidelity emulation and lower fidelity simulation to facilitate RL for the coevolution of red and blue agents capable of executing cyber operations."[76] By combining simulations (to train agents) and emulations (to adapt these to 'real life' scenarios), these are important developments that may offer a foundation to recursively enable AI systems to mount cyber-attacks in the future. This could be done by exploiting certain vulnerabilities, and other potential attack vectors to enable greater offensive cyber operations that permit operational engagement either directly, or by informing new methods that may be used.

Some further applications could include the adaption of deep RL to conventional 'penetration' testing[77] to identify vulnerabilities in computational systems. Although this could benefit defence, it could permit offensive pivots on networks, such as through the tool DeepExploit,[78] or to even exploit different elements of the cyber 'kill chain.'[79] This could have significant effects on cyber operations, which could allow bespoke, highly obfuscated attacks from unusual vectors, or to generate new perspectives on vulnerabilities. Yet, as is common with other applications of deep RL, there are significant issues on the complexity of computational systems and their ecologies. Deep RL arguably optimises better in 'data rich' environments, so as states such as the US move towards persistent engagement and aim to employ stealthy measures to gain strategic advantage,[80] the risk of using malware with deep RL may be too risky, for example. Even slight ecological changes for a deep RL agent could be troublesome, which may lead to its application in more, perhaps 'indiscriminate,' widespread attacks which are not dependent on stealth – meaning its impact on state offensive cyber operations will remain inconclusive. This is particularly so as the ethics of offensive cyber increasingly considers the impact of AI on operations.[81] Questions will also be raised on how this could fit into broader ethical, legal and operational aspects of emerging state organisations for offensive cyber operations, such as the UK's National Cyber Force.[82]

Environments

Deep RL has the capacity to be applied to a range of conflict scenarios, such as those highlighted above. Although advances have been made since *AlphaGo Zero* in 2017,[83] such as with *AlphaStar* discussed above, there are persistent issues of data availability when in conflict situations and complex environmental and ecological dynamics. Yet as Lee et al. note, even in cases where there is a partiality of data and knowledge of the environment, agents performed better than humans in *StarCraft II*.[84] Regardless, deep RL is still primarily suited to 'structured problems' which allow for an optimal strategy to be converged upon for full game simulations. That is, adaptions to games

will not easily lend themselves to real-world adversarial engagements.[85] As is reflected on reports of the promise of AI to cyber conflict, (deep) RL is not likely to be useful in the immediate term.[86] So, the rest of this chapter reflects on how deep RL could contribute to a changing dynamic of conflict, where decisions may *still* appear to be made by humans but which are actually being built in co-constitutive ways.[87]

Although the work on *StarCraft II* by DeepMind is a "well recognized new milestone for AI research, [*StarCraft II*] continues to present a grand challenge to deep RL due to its complex visual input, large action space, imperfect information, and long horizon."[88] However, even more recent developments by DeepMind on *MuZero* combine different algorithms into a broader AI system that "does not require any knowledge of the environment dynamics, potentially paving the way towards the application of powerful learning and planning methods to a host of real-world domains from which there exists no perfect simulator."[89] This demonstrates the fast pace of change, and for how militaries and others may quickly pivot and adapt to the potential of these technologies.

Environments will remain an essential and limiting factor to their deployment, in understanding their dynamics, and how 'agents' will work within them. Any computational representation of an environment is an abstraction, and thus there is always within it the potential of surprise and difference. Within conflict, environments are crucial, so that we could say that deep RL is not just about agents and adversarial behaviours but also a new form of battlespace, a space that retreats, and reconstitutes conflict itself. In so doing, the security of such environments is paramount, as they are likely to be targets if they are used for strategy or even offensive cyber operations.[90] Such environments could be subtly manipulated by adversarial inputs that are imperceptible to humans, where computational recognition is changed just enough to transform the outcome of a strategy or operation beyond a human capacity to intervene. This can include adversarial attempts against deep RL with limited knowledge (black box) to those with complete knowledge (white box) with a variety of different methods, such as data poisoning.[91] Thus, here we can see that deep RL will not 'solve' conflict but may appear to offer new perspectives where agents act through approximation and optimisation, introducing greater risks to the 'control' of conflict. Such attacks, where incommensurability is a condition of its performance, then add yet another *imperceptible trace* of unknowability in their ethico-political standing.

Tracing conflict

> The emergence of technological autonomy linked to advances in AI and represented by the growing research interest in algorithms introduces a novel, potentially game-changing element for the mature models of human agency and material/nonmaterial structure dominant in IR theory.[92]

As Hendrik Huelss reflects on norms and algorithms in international relations, AI systems have the capacity to radically challenge how human agency and computation come together. As computation has a capacity for choice-making, this challenges conventional technological thinking as it actively constructs what is optimal and how conflict may be practiced, albeit intersected with human decision-making, as a specific form of reflexive choice. Yet, as I have hinted towards, there are imperceptible connections, or traces, that lead to new formulations when conflict is unknowable in relation to deep RL. In this section, I shall outline how traces could help understand the complex interplay across incommensurability for both offensive cyber operations and strategic wargaming.

AI systems are reliant on big data to generate their own 'ground truths,' but this is often hosted through cloud computing servers, distributed and distanced from the 'face' of conflict. Conflict has conventionally occurred *over there*, but as with other forms of security these have been brought into the everyday, where processing of data and actions are made at great distances from where they may be performed, such as in drone warfare.[93] With deep RL, it is not only about *where* cloud computing and algorithmic processing are occurring but also that new spaces and notions of conflict are being created *by* and *within* deep RL. By this, I mean the data and assumptions for environments, the possible actions of an agent, and their optimised and approximated rewards, are articulating new worlds and possibilities. International conflict, as such, could be actively performed in the spaces of the adversarial relationships of deep RL, where computational choices, human decisions, designers, environments, big data and more come together to optimise the future. Some of these futures may become understood as more probable, more likely, but only in conversation with these others. These transformations could lead to more subtle changes in the norms of international conflict, and its traces far more complex than could be immediately imagined.

One method to exemplify tracing conflict is turning attention to the neurons that form the 'deep' part of deep RL, the artificial neural networks. Some methods are seeking to expose how to 'reveal' how a neural network may have 'seen' features, especially in image recognition.[94] However, I argue, along with Fazi,[95] that there is an incommensurability here, so that even if values partially represent choices made by computation, there will always be an unknowability beyond representational critique, a trace of another recognition. In simpler terms, there is no adequate translation available from deep RL to human social negotiation or norms. In Figure 2.1, I present a very abstracted (cybernetic and neuroscience inspired) representation of a neuron as is common in computer science literature. On the left, arrows point towards the circle. The arrows each have a weight w_n that are calculated with a 'learnt' bias Σ that then produces an output y_1. Input weights may emerge from previous neurons in neural layers, that each output an 'activation' with its own weight to pass on to the next layer. There are many different variations of this process, which 'layer' it is in, and whether certain neurons are activated at all.

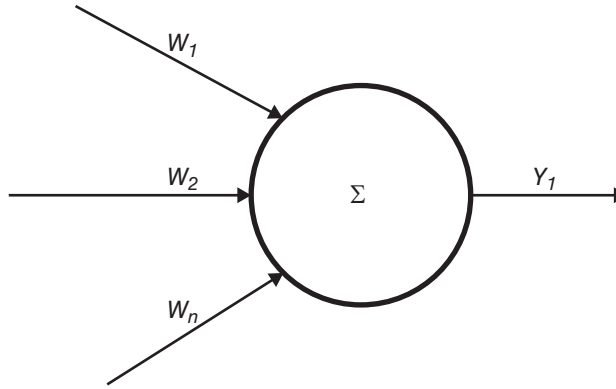

Figure 2.1 A simplified diagrammatic of an artificial neural network algorithm neuron, as popularised through cybernetic comparison.
Source: Author's creation.

Here, we can sense the traces of such a process, but never intuitively feel or comprehend wholly what these may be.

Even if multiple options are presented in the output of the neural network algorithm for the reward value, one cannot wholly excavate *why* such an outcome of the loss function was made. It may be possible to look at the activation points of a neuron, but it would be difficult to assess how this then comes to interact across the environment, the agent, and the policy – where computational choices intersect with one another so that it may be become unclear what is exactly leading to a strategic offer or the best method to target an adversary. This is only magnified when there are multiple systems. Crucially, neurons operate to condense possibility and must have bias (intended or unintended) to function.[96] Thus, they must condense all the possibility, and in the case of deep RL, approximate a reward. To trace conflict is not to identify points of activation but to incorporate contexts from big data, their biases, and how these may be used in determining an environment for the agent. If data is used from datasets outside previous spaces of conflict, then conflict becomes intertwined with these, where they become unknowingly implicated into the practice of conflict itself.

Tracing conflict is not just about the isolation of neurons within deep RL – or other AI systems – away from the big data, environments, and ecologies we live in, but the way AI systems wrap up contexts and previous weights of human decisions and computational choices in complex hybrids. This argument can work alongside the good work that is ongoing to identify biases within learning data or to show potential points. However, this will always be insufficient to address the complexity of the traces that emerge in assessing conflict. This is one of the trickiest of puzzles. It is not only that

there is not *enough* data, particularly in conflict situations, for there to be robust assessments of how AI will react[97] but also that the greater the volume of data, the more unknown conflict becomes. Thus, both too much and too little data, amid the impossibility of ever truly 'simulating' an environment, suggests a persistent unknowability that emerges with AI systems.

Therefore, to even consider deploying highly recursive systems is not about the 'events' or points of decision of the human in the loop, but rather about how to incorporate the doubt of this computational unknowability, and to assess if it is sensible to involve it in conflict at all. A neural network must condense contexts of the past to produce a future output. AI systems collapse different places, contexts, and knowledge into a particular moment where weights are produced to suggest the significance of this or that action. Future conflicts with deep RL will be both hyper-local to the learning algorithms as much as they are yet distributed simultaneously across learning data and simulated environments, bending our notions of scale and even the international through their traces. When conflict is mediated through computational forms of recognition to identify an optimum strategic output, the political space for failure, chances for social reflection recede. Where is doubt? If actions by deep RL suggest a future action that cannot be understood now, should that ever be taken? These are crucial ethical and political questions that cannot be solved through calculative forms of recognition. As Jacques Derrida contends, decisions are not of the order of the calculable.[98] But decision cannot be 'inserted' into the machine learning algorithm as ethics precedes a codification. Nor can its outputs be recaptured by *the loop* of the human to somehow reassert authority and control, as computational recognition disturbs the knowability and commensurability of the outputs themselves.

Trace as an ethics beyond 'the loop'

We might understand that deep RL *performs* in multiple spaces and across contexts, even if it is *located* on a particularly 'secure' server within the UK or US for instance. Therefore, if multiple militaries adopt deep RL, it is possible to understand this as conflict retreating to forms of computation cognition and choice. This may appear to be a radical claim, but Chinese strategic thinking of an AI singularity,[99] as well as a common 'winner-takes-all' approach that sustains military thinking, point towards this as one potential trajectory. Yet, in such practices, conflict conducted by computation should not be considered as conflict, when conflict is informed by a humanly impulse, emotion, and social sensibility. I take such a position from thinking with Emmanuel Levinas, and his work on the notion of the face.[100] That is, our intrinsic human condition is to be in relation to the other before we can constitute ourselves, that has also been used by Judith Butler in feminist thinking on violence.[101] Although there is not space here to go into the details behind such thinking, the importance is that, for Levinas, ethics arrives before its articulation, in

relation with the *unknown* other.[102] In simpler terms, ethics is not something that comes after ourselves, but is inherent to us. Ethics cannot thus be codified and inserted into deep RL. In offensive cyber operations, if the deep RL simply optimises the output with steps that may be unknowable, it is not conducting this through a social understanding of norms or ethics. The face is not a physical manifestation but affords a capacity for an ethics before it has been witnessed the other. When understanding computation, it does not have a social recognition of the face, and thus it is indifferent and incapable of ever attaining a social ethics of the other.

A trace according to Jacques Derrida is "impossible-unthinkable-unsayable"[103] in that it connects us in ways that are not directly *traceable*. I use this in my reading of the potential for AI systems to conduct conflict to claim that there will never be a wholly 'explainable AI' of its multiple traces as much as it cannot fully bear the weight of contexts past. In this chapter, then, I do not seek to see the trace as something which can be explicitly drawn. This is a trace that is unable to be wholly articulated, it is incommensurable. However, traces hold us accountable to other, to the face, in our social negotiation of norms and ethics. The reason to raise such a point is how the *face* of conflict is incommensurable to computation, even as many are frantically trying to cement an authoritative human in the loop to stamp an anthropomorphic face on deep RL and other AI systems. In wargaming, the trace of the human other through our social negotiation is always there, but it is incommensurable to deep RL. Although, undoubtedly, there are traces of *our* ethics and politics, they get twisted and contorted to the point that our own social negotiation may become incommensurable to itself: that is, our ethics morphs towards optimisation and means that conflict itself becomes a zero-sum goal.

As I have said, AI systems do not abstract in a human sensibility of our social connections, meaning that my claim is that it is not sensible to consider that computation generates a similar ethico-political construct of conflict, as much as it abstracts away from, and sterilises the possibility of such conflict. As Bergen and Verbeek state, Levinas did not explicitly address technology, but they suggest that technology with a 'face' has emerged as they have become 'humanized.'[104] With AI systems, this is most clear in the cybernetic comparison to human intelligence, as if it is on an equivalent plane to computational recognition and choice. This is not the case, as computation is not necessarily on a hierarchy above or beneath us but offers a different way of being in the world. The danger is that deep RL is understood to have a face through its interfaces and outputs, anthropomorphising computation with a trust roughly equivalent to a human ethico-political sensibility.

So, why focus on such a point of the face for the ethico-political future of conflict? As recognition is increasingly performed by computation, its calculative logics and choices will become key to articulate the norms and practice of conflict. Thus, in a deep RL, the 'agent' does not embody the face of the other in its recognition. This is because it does not recognise or relate on an equivalent basis to us. It is incommensurable to our ethico-political

frameworks. It is not only in LAWS that we must focus our attention but also on offensive cyber operations, wargaming, and elsewhere that are not *immediately* lethal and on the 'front' line. AI systems will condition organisations and bureaucracies to plan and operationalise future conflict. It may be that there is healthy scepticism of such outcomes of deep RL in military planning, but the solution is not to have *the loop*. As Schwarz details, "the role of the human as a moral decision-maker in the technological loop is considerably more complex than the picture of the functional and rational agent 'in' or 'on' the loop might suggest."[105] Deep RL emerges as a product of multiple authors – software engineers, learning data, assumptions, decision-makers priorities, computational recognition and choice, and more – thus rather than conceptualising *the loop*, humans and others are already intrinsically, unknowably, enmeshed within it through traces. This changes the ethico-political relation, when the adversary is no longer simply face to face but rather becomes assumed, constructed, and abstracted and partially recognised through traces by computation.

Traces expose the falsity of *the loop*. So, when it is said 'is the human there,' can that human have a 'meaningful' say, is to assume that there is a possibility for us to translate the face through computation. As conflict is mediated and interpreted by AI systems, conflict ceases to be conflict in a meaningful sense. Instead, rather than understanding AI systems and conflict as something that can regarded as 'evolutionary' or 'revolutionary' in debates on technological change in international relations,[106] conflict is beyond, *inter alia,* human knowability, when AI systems grow in application. If it down to approximation, calculability, and unknowability, then conflict is transformed, even if there is an ability to say 'no' to this or that particular action – such as *that* cyber-attack or *this* plan emerging from a wargame.

Conclusions: unknowability and persistence

Warfare and conflict are not only unknowable due to a lack of data or the 'fog of war' but also due to the face that is incommensurable to AI systems as well as inadequate abstractions and translations of outputs made though computational recognition and choice for humans. In this chapter, I have offered both a conceptual and practical reflection on deep RL – through wargaming and offensive cyber operations – to detail both a *general* transformation of conflict as well as the *particular* of the difficulties of operationalising such technologies. There is no possibility of an escape from traces in conflict, but computation has seriously complicated how these traces lead to an ethico-political sensibility. Bellanova et al. argue that there is a "force of computation"[107] that perpetuates certain forms of violence and harm, that I think can be applied to the incommensurability of a face to AI systems. Across 'cognitive assemblages,' computation's calculative recursivity and recognition will transform how militaries, governments, and corporations understand conflict through AI systems. As there is no possibility to codify computation with human

sensibilities, this is particularly dangerous, not only for lethal autonomous weapons but also for the broader *optimisation* of conflict in the future.

I have sought to trace how deep RL transforms how international conflict may be understood. First, I have considered *where* conflict may be conducted and have argued that, through deep RL, it will occur through new battlespaces that occur both through and *within* AI systems. This transforms what is considered as the spaces of international conflict as multiple deep RL agents simulate the adversary, conflict is no longer just on the 'frontline,' but is withdrawn and retracted, disrupting what the 'international' of conflict could be. This is amplified by traces that becomes threaded across AI systems and humans, which expand the sources of what informs the practice of conflict. Second, I have assessed *how* conflict is transforming through computational cognition and choice. This has demonstrated how recursivity has been central to transforming dynamics of control, away from a deterministic tool, to one that actively participates in the performance of conflict. In offensive cyber operations, this could be 'agents' deployed on an adversary network, to suggested strategic moves in wargaming that inform how a military or government may respond. And third, I have questioned *who* is performing conflict. As computational recognition and choice challenge the commensurability between human and social norms on the one hand and calculative norms on the other, conflict becomes distorted. As I have explored through the face with Levinas, the norms of what is considered ethical or appropriate, will be expressed with and through AI systems. This means that the social norms of international conflict could be slowly transformed. Thus, AI systems introduce an unknowability to international conflict and its performance through cyberspace. That is, it is unknowable exactly (from) where conflict is taking place, how conflict is being performed and constructed, and who (and what) is doing such work.

Ultimately, there is no 'outside' for a rational human in, on, or out of *the loop*, as we are always imbricated within it. It is not that there is a computational choice and human decision maker that exist in a binary divide. Instead, they are co-constitutive, part of ecologies of interaction of organisational cultures, computational choices, decisions, data, and so on. It is recursivity that has become so important – even as computation has been key to the enactment of international conflict for many decades. This is what makes conflict and AI systems profoundly different, there is no outside with which to retrieve an ethics as much as there is no ability to wholly steer it from within; AI systems have no one person who has the capacity to know all the traces and can exert their explicit intent. Yet more insidiously than the control promised by *the loop*, computation transforms norms through its outputs. Therefore, in embracing this unknowability, I claim that there is a danger that in the hope of deploying operations with speed and new insights offered by computational recognition, there is a loss of the ethical and political face of conflict, where optimisation holds sway. So, in advocating for the death of the loop, I hope to prevent the slow death of our ethico-political relationship to one another.

Notes

1 Hendrik Huelss, "Norms are what machines make of them: Autonomous weapons systems and the normative implications of human-machine interactions," *International Political Sociology* 14, no. 2 (2020).
2 What may be considered 'undesirable' may be hotly contested. However, computation is still a logical system and thus must 'execute' what is programmed. How this intersects with particular ecologies (see, for example, my previous work in Andrew Dwyer, "Malware ecologies: A politics of cybersecurity," PhD Thesis, University of Oxford, 2019) is of crucial importance to whether something may be desirable or not, to particular people, in certain places.
3 I refer to AI systems to express that AI is often a complex mix of components consisting of software, machine learning algorithms, data, people, and organisations that come together as a broader collective that can be understood as a 'system.'
4 Elke Schwarz, "Autonomous weapons systems, artificial intelligence, and the problem of meaningful human control," *The Philosophical Journal of Conflict and Violence* 5, no. 1 (2021).
5 Thilo Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *Minds and Machines* 30, no. 1 (2020); Jessica Morley et al., "Ethics as a service: A pragmatic operationalisation of AI ethics," *Minds and Machines* 31 (June 2021).
6 Peter Asaro, "On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making," *International Review of the Red Cross* 94, no. 886 (2012); Ingvild Bode and Hendrik Huelss, "Autonomous weapons systems and changing norms in international relations," *Review of International Studies* 44, no. 3 (2018).
7 Beatrice M. Fazi, "Beyond human: Deep learning, explainability and representation," *Theory, Culture & Society* 38, no. 7–8 (November 2020); Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GPDR," *Harvard Journal of Law & Technology* 31, no. 2 (2017).
8 In September 2020 ransomware affected computational systems at a hospital Düsseldorf, Germany. This was proclaimed as the first death caused by a cyberattack. However, it is highly disputed whether this was the case (see William Ralston, "The untold story of a cyberattack, a hospital and a dying woman." *WIRED*, 11 November 2020).
9 Andrew Dwyer, "Cybersecurity's grammars: A more-than-human geopolitics of computation," *Area*, Online First (2021).
10 Benjamin M. Jensen, Christopher Whyte, and Scott Cuomo, "Algorithms at war: The promise, peril, and limits of artificial intelligence," *International Studies Review* 22, no. 3 (2020).
11 Louise Amoore, *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*, (Durham, NC: Duke University Press, 2020).
12 Michel Foucault, *Society Must Be Defended: Lectures at the Collège de France, 1975–1976* (London: Penguin Books, 2003), 15.
13 Tim Stevens, "Knowledge in the grey zone: AI and cybersecurity," *Digital War* 1 (March 2020).
14 See Yuk Hui, *Recursivity and Contingency* (London: Rowman & Littlefield International, 2019).
15 Schwarz, "Autonomous weapons systems;" Katherine N. Hayles, *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics* (Chicago: University of Chicago Press, 1999).
16 Hui, *Recursivity*.
17 Beatrice M. Fazi *Contingent Computation: Abstraction, Experience, and Indeterminacy in Computational Aesthetics* (London: Rowman & Littlefield International, 2018); Dwyer, "Cybersecurity's grammars."

18 Although see for this discussion: Vincent Boulanin and Maaike Verbruggen, "*Mapping the Development of Autonomy in Weapon Systems* (Stockholm, Sweden: Stockholm International Peace Research Institute, 2017).
19 Amanda Sharkey, "Autonomous weapons systems, killer robots and human dignity," *Ethics and Information Technology* 21, no. 2 (2019); Schwarz, "Autonomous weapons systems."
20 Development, Concepts and Doctrine Centre. "Human-machine teaming." *Joint Concept Note 1/18*. Swindon: Ministry of Defence, 2018.
21 Jensen, Whyte and Cuomo, "Algorithms at war."
22 Schwarz, "Autonomous weapons systems."
23 Lipika Majumdar Roy Choudhury et al., "*Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973.* (New York: United Nations Security Council, 2021).
24 Amoore, *Cloud Ethics*, 5.
25 This is not to claim that previous technologies have not had great affordance and impact, and completely transformed societies and forms of conflict. However, none have had a recursive capacity in the same way. Whilst there may be humans behind each algorithm, they also do not have the same ethical and political relations as in the past.
26 Keith Dear, "Artificial intelligence and decision-making," *The RUSI Journal* 164, no. 5–6 (2019): 25.
27 Jussi Parikka, "Ethologies of software art: What can a digital body of code do?" in *Deleuze and Contemporary Art* (Edinburgh: Edinburgh University Press, 2010).
28 Louise Amoore and Rita Raley, "Securing with algorithms: Knowledge, decision, sovereignty," *Security Dialogue* 48, no. 1 (2017): 4.
29 Casey R. Lynch and Vincent J. Del Casino, "Smart spaces, information processing, and the question of intelligence," *Annals of the American Association of Geographers* 110, no. 2 (2020): 2.
30 For the issues that occur in such comparison, see: Johannes Bruder, *Cognitive Code: Post-Anthropocentric Intelligence and the Infrastructural Brain* (Montreal: McGill-Queen's University Press, 2019).
31 Here I refer to Turing machines, electronic digital binary computation, rather than emerging forms of technology, such as quantum computing which could offer different potentials (Vedran Dunjko, Jacob M. Taylor, and Hans J. Briegel, "Quantum-enhanced machine learning," *Physical Review Letters* 117, no. 130501 (2016)).
32 Katherine N. Hayles, *Unthought: The Power of the Cognitive Nonconscious* (Chicago: University of Chicago Press, 2017).
33 Luciana Parisi, "Critical computation: Digital automata and general artificial thinking," *Theory, Culture & Society* 36, no. 2 (2019): 109.
34 Fazi, "Beyond human."
35 Amoore, *Cloud Ethics*, 16.
36 Lucy Suchman, Karolina Follis, and Jutta Weber, "Tracking and targeting: Sociotechnologies of (in) security," *Science, Technology, & Human Values* 42, no. 6 (2017).
37 Adrian Mackenzie, *Machine Learners: Archaeology of a Data Practice* (Cambridge, MA: MIT Press, 2017).
38 Amoore, *Cloud Ethics*, 65.
39 Schwarz, "Autonomous weapons systems," 57.
40 Tim Stevens, *Cyber Security and the Politics of Time* (Cambridge: Cambridge University Press, 2016), 93.
41 Jensen, Whyte and Cuomo, "Algorithms at war."
42 For an accessible and easy to read blog on this distinction, see Ariel Liu, "An introduction to machine learning," *DataSeries* (blog on Medium), 16 September 2019.

43 Yann Le Cun, *Quand La Machine Apprend: La Révolution Des Neurones Artificiels et de l'apprentissage Profond* (Paris: Odile Jacob, 2019).
44 Hui, *Recursivity*, 114–115.
45 David Silver and Demis Hassabis, "AlphaGo zero: Starting from scratch," *DeepMind* (blog), 18 October 2017.
46 Volodymyr Mnih et al., "Human-level control through deep reinforcement learning," *Nature* 518, no. 7540 (2015).
47 Convolutional neural network algorithms are used for image recognition.
48 Richard S. Sutton and Andrew G Barto, *Reinforcement Learning: An Introduction* (Cambridge, MA: MIT Press, 2018), 2.
49 Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning. Adaptive Computation and Machine Learning* (Cambridge, MA: MIT Press, 2016).
50 Dennis Lee et al., "Modular architecture for starcraft II with deep reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (2018), 188.
51 Mnih, "Human-level control," 529.
52 David Silver et al., "Reward is enough," *Artificial Intelligence* 299 (October 2021): 10.
53 Marc W. Brittain and Peng Wei, "One to any: Distributed conflict resolution with deep multi-agent reinforcement learning and long short-term memory," in *AIAA Scitech 2021 Forum* (New York, N.Y: American Institute of Aeronautics and Astronautics, 2021).
54 Revanth Konda, Hung Manh La, and Jun Zhang, "Decentralized function approximated Q-learning in multi-robot systems for predator avoidance," *IEEE Robotics and Automation Letters* 5, no. 4 (2020).
55 Hassam Ullah Sheikh, Mina Razghandi, and Ladislau Bölöni, "Learning distributed cooperative policies for security games via deep reinforcement learning," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)* (IEEE, 2019).
56 Roger Smith, "The long history of gaming in military training," *Simulation & Gaming* 41, no. 1 (2010).
57 Andreas Haggman, "Cyber wargaming: Finding, designing, and playing wargames for cyber security education," PhD Thesis, Royal Holloway, University of London, 2019; Benjamin Schechter, Jacquelyn Schneider, and Rachael Shaffer, "Wargaming as a methodology: The international crisis wargame and experimental wargaming," *Simulation & Gaming* 52, no. 4 (2021); Fabio Cristiano, "From simulations to simulacra of war: Game scenarios in cyberwar exercises," *Journal of War & Culture Studies* 11, no. 1 (2018).
58 James Goodman, Sebastian Risi, and Simon Lucas, "AI and Wargaming," arXiv:2009.08922 [cs.AI] (2020).
59 Elsa B. Kania, "*Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power*" (Center for a New American Security, 2017), 13.
60 Agent-based models simulate the actions of autonomous agents, enabling to understand thebehaviour of a system. This is distinct to deep RL, which recursively iterates adversary behaviours and how these may change the environment and their connection to one another.
61 Ministry of Defence, "*Digital Strategy for Defence: Delivering the Digital Backbone and Unleashing the Power of Defence's Data*" (London: UK Government, 2021).
62 Daniel Griffiths, "Improbable's defence business awarded contract for a second year," *Improbable*, 12 November 2020.
63 Hanchao Wang, "Large scale deep reinforcement learning in war-games," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE, 2020).
64 Goodman, Risi and Lucas, "AI and wargaming."
65 Ibid., 15.

66 "Les simulateurs doivent être suffisamment puissants et précis, c'est-à-dire refléter assez précisément ce qui se passe dans la réalité pour que, une fois le système entraîné par simulation, on puisse transposer sa capacité au monde réel. Cela n'est pas toujours possible." Author translation of Le Cun, *Quand La Machine*, 303.
67 "StarCraft is a real-time strategy game in which players balance high-level economic decisions with individual control of hundreds of units. This domain raises important game-theoretic challenges: it features a vast space of cyclic, non-transitive strategies and counterstrategies; discovering novel strategies is intractable with naive self-play exploration methods; and those strategies may not be effective when deployed in real-world play with humans. Furthermore, StarCraft has a combinatorial action space, a planning horizon that extends over thousands of real-time decisions, and imperfect information." (Oriol Vinyals et al., "Grandmaster level in starcraft II using multi-agent reinforcement learning," *Nature* 575, no. 7782 (2019): 350.)
68 Bounded state refers to those problems which have a finite number of moves, which deep RL was initially used in. Therefore, the most famous advances were against games with set rules and spaces, such as *Go*.
69 Vinyals, "Grandmaster level," 353.
70 Goodman, Risi and Lucas, "AI and wargaming," 43.
71 Mary Cummings, "Automation bias in intelligent time critical decision support systems," in *AIAA 1st Intelligent Systems Technical Conference*. (American Institute of Aeronautics and Astronautics, 2004).
72 This is sometimes referred to as the 'fog of war' or more simply as places where little data is available.
73 Ben Buchanan et al., "*Automating Cyber Attacks: Hype and Reality*" (Center for Security and Emerging Technology, 2020), 21.
74 Thanh Thi Nguyen and Vijay Janapa Reddi, "*Deep Reinforcement Learning for Cyber Security,*" arXiv:1906.05799 [cs.CR] (2019).
75 Callum Baillie et al., "*Cyborg: An Autonomous Cyber Operations Research Gym,*" arXiv:2002.10667 [cs.CR] (2020).
76 Ibid., 4.
77 László Erdődi and Fabio Massimo Zennaro, "The agent web model: Modeling web hacking for reinforcement learning," *International Journal of Information Security* 21 (June 2021); Jonathon Schwartz and Hanna Kurniawati, "*Autonomous Penetration Testing Using Reinforcement Learning*," arXiv:1905.05965 [cs.CR] (2019).
78 Isao Takaesu, "DeepExploit," *Github*, 2018.
79 Buchanan et al., "Automating cyber attacks."
80 Max Smeets, "US cyber strategy of persistent engagement & defend forward: implications for the alliance and intelligence collection," *Intelligence and National Security* 35 no. 3 (2020).
81 Ewan Lawson and Kubo Mačák, "*Avoiding Civilian Harm from Military Cyber Operations During Armed Conflicts*," (Geneva, Switzerland: International Committee of the Red Cross, 2021).
82 Joe Devanny et al., "*The National Cyber Force That Britain Needs?*" (London: King's College London, 2021).
83 Silver and Hassabis, "AlphaGo zero."
84 Lee et al., "Modular architecture."
85 George Cybenko, "Adaptation and deception in adversarial cyber operations," in *Modeling and Design of Secure Internet of Things* (Piscataway, NJ, USA: IEEE Press, 2020).
86 Wyatt Hoffman, "*AI and the Future of Cyber Competition*." CSET Issue Brief (Center for Security and Emerging Technology, 2021); Buchanan et al., "Automating cyber attacks."

87 Mackenzie, *Machine Learners*.
88 Lee et al., "Modular architecture," 187.
89 Julian Schrittwieser et al., "Mastering atari, go, chess and shogi by planning with a learned model," *Nature* 588 (2020): 608.
90 Matteo E. Bonfanti, Myriam Dunn Cavelty, and Andreas Wenger, "Artificial intelligence and cyber-security," in *The Routledge Social Science Handbook of AI* (London: Routledge, 2021).
91 Vasisht Duddu, "A survey of adversarial machine learning in cyber warfare," *Defence Science Journal* 68, no. 4 (2018).
92 Huelss, "Norms are what machines," 7.
93 Derek Gregory, "From a view to a kill: drones and late modern war," *Theory, Culture & Society* 28, no. 7–8 (2011).
94 Alejandro Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion* 58 (June 2020).
95 Fazi, "Beyond human."
96 Amoore, *Cloud Ethics*.
97 Jensen, Whyte and Cuomo, "Algorithms at war."
98 Jacques Derrida, *Of Grammatology* (Baltimore: Johns Hopkins University Press, 2016).
99 Kania, "Battlefield Singularity."
100 Emmanuel Levinas, *Totality and Infinity: An Essay on Exteriority* (The Hague: Martinus Nijhoff Publishers, 1979).
101 Judith P. Butler, *Giving an Account of Oneself* (New York: Fordham University Press, 2005); *Precarious Life: The Powers of Mourning and Violence* (London: Verso, 2006).
102 Caroline Holmqvist, "Undoing war: War ontologies and the materiality of drone warfare," *Millennium* 41, no. 3 (2013).
103 Robert Bernasconi, "The trace of levinas in derrida," in *Derrida and Différance* (Evanston, IL: Northwestern University Press, 1988), 17.
104 Jan Peter Bergen and Peter-Paul Verbeek, "To-do is to be: Foucault, levinas, and technologically mediated subjectivation," *Philosophy & Technology* 34 (2021).
105 Schwarz, "Autonomous weapons systems," 62.
106 Jensen, Whyte and Cuomo, "Algorithms at war."
107 Rocco Bellanova et al., "Toward a critique of algorithmic violence," *International Political Sociology* 15, no. 1 (2021): 123.

Bibliography

Amoore, Louise. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham, NC: Duke University Press, 2020.

Amoore, Louise, and Rita Raley. "Securing with algorithms: Knowledge, decision, sovereignty." *Security Dialogue* 48, no. 1 (2017): 3–10. https://doi.org/10.1177/0967010616680753.

Asaro, Peter. "On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making." *International Review of the Red Cross* 94, no. 886 (2012): 687–709. https://doi.org/10.1017/S1816383112000768.

Baillie, Callum, Maxwell Standen, Jonathon Schwartz, Michael Docking, David Bowman, and Junae Kim. "*Cyborg: An Autonomous Cyber Operations Research Gym*." arXiv:2002.10667 [cs.CR] (2020).

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia. "Explainable artificial

intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* 58 (June 2020): 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Bellanova, Rocco, Kristina Irion, Katja Lindskov Jacobsen, Francesco Ragazzi, Rune Saugmann, and Lucy Suchman. "Toward a critique of algorithmic violence." *International Political Sociology* 15, no. 1 (2021): 121–150. https://doi.org/10.1093/ips/olab003.

Bergen, Jan Peter, and Peter-Paul Verbeek. "To-do is to be: Foucault, levinas, and technologically mediated subjectivation." *Philosophy & Technology* 34 (2021): 325–348. https://doi.org/10.1007/s13347-019-00390-7.

Bernasconi, Robert. "The trace of levinas in derrida." In *Derrida and Différence*, edited by David Wood and Robert Bernasconi, 13–30. Evanston, IL: Northwestern University Press, 1988.

Bode, Ingvild, and Hendrik Huelss. "Autonomous weapons systems and changing norms in international relations." *Review of International Studies* 44, no. 3 (2018): 393–413. https://doi.org/10.1017/S0260210517000614.

Bonfanti, Matteo E., Myriam Dunn Cavelty, and Andreas Wenger. "Artificial intelligence and cyber-security." In *The Routledge Social Science Handbook of AI*, edited by Anthony Elliott, 222–236. London: Routledge, 2021. https://doi.org/10.4324/9780429198533.

Boulanin, Vincent, and Maaike Verbruggen. "*Mapping the Development of Autonomy in Weapon Systems*." Stockholm, Sweden: Stockholm International Peace Research Institute, 2017. https://web.archive.org/web/20210912125048/https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf.

Brittain, Marc W., and Peng Wei. "One to any: Distributed conflict resolution with deep multi-agent reinforcement learning and long short-term memory." In *AIAA Scitech 2021 Forum*. New York, NY: American Institute of Aeronautics and Astronautics, 2021. https://doi.org/10.2514/6.2021-1952.

Bruder, Johannes. *Cognitive Code: Post-Anthropocentric Intelligence and the Infrastructural Brain*. Montreal: McGill-Queen's University Press, 2019.

Buchanan, Ben, John Bansemer, Dakota Cary, Jack Lucas, and Micah Musser. "*Automating Cyber Attacks: Hype and Reality*." Center for Security and Emerging Technology, 2020. https://web.archive.org/web/20210317113544/https://cset.georgetown.edu/wp-content/uploads/CSET-Automating-Cyber-Attacks.pdf.

Butler, Judith P. *Giving an Account of Oneself*. New York: Fordham University Press, 2005.

Butler, Judith P. *Precarious Life: The Powers of Mourning and Violence*. London: Verso, 2006.

Cristiano, Fabio. "From simulations to simulacra of war: Game scenarios in cyberwar exercises." *Journal of War & Culture Studies* 11, no. 1 (2018): 22–37. https://doi.org/10.1080/17526272.2017.1416761.

Cummings, Mary. "Automation bias in intelligent time critical decision support systems." In *AIAA 1st Intelligent Systems Technical Conference*. American Institute of Aeronautics and Astronautics, 2004.

Cybenko, George. "Adaptation and deception in adversarial cyber operations." In *Modeling and Design of Secure Internet of Things*, edited by Charles A. Kamhoua, Laurent L. Njilla, Alexander Kott, and Sachin Shetty, 111–22. Piscataway, NJ: IEEE Press, 2020. https://doi.org/10.1002/9781119593386.ch5.

Dear, Keith. "Artificial intelligence and decision-making." *The RUSI Journal* 164, no. 5–6 (2019): 18–25. https://doi.org/10.1080/03071847.2019.1693801.

Derrida, Jacques. *Of Grammatology*. Translated by Gayatri Chakravorty Spivak. Baltimore: Johns Hopkins University Press, 2016.

Devanny, Joe, Andrew Dwyer, Amy Ertan, and Tim Stevens. "*The National Cyber Force That Britain Needs*?" London: King's College London, 2021. https://web.archive.org/web/20210917142536/https://www.kcl.ac.uk/policy-institute/assets/the-national-cyber-force-that-britain-needs.pdf.

Development, Concepts and Doctrine Centre. "*Human-Machine Teaming*." Swindon: Ministry of Defence, 2018. https://web.archive.org/web/20210917142447/https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/709359/20180517-concepts_uk_human_machine_teaming_jcn_1_18.pdf.

Duddu, Vasisht. "A survey of adversarial machine learning in cyber warfare." *Defence Science Journal* 68, no. 4 (2018): 356–366. https://doi.org/10.14429/dsj.68.12731.

Dunjko, Vedran, Jacob M. Taylor, and Hans J. Briegel. "Quantum-enhanced machine learning." *Physical Review Letters* 117, no. 130501 (2016). https://doi.org/10.1103/PhysRevLett.117.130501.

Dwyer, Andrew. "*Malware Ecologies: A Politics of Cybersecurity*." PhD Thesis, University of Oxford, 2019. https://web.archive.org/web/20210917142218/https://ora.ox.ac.uk/objects/uuid:a81dcaae-585b-4d5b-922f-8c972b371ec8/download_file?file_format=pdf&safe_filename=Malware_Ecologies_Dwyer_Bodleian_Copy.pdf&type_of_work=Thesis.

Dwyer, Andrew. "Cybersecurity's grammars: A more-than-human geopolitics of computation." *Area*, Online First (2021). https://doi.org/10.1111/area.12728.

Erdődi, László, and Fabio Massimo Zennaro. "The agent web model: modeling web hacking for reinforcement learning." *International Journal of Information Security* 21 (June 2021): 293–309.

Fazi, M. Beatrice. *Contingent Computation: Abstraction, Experience, and Indeterminacy in Computational Aesthetics*. London: Rowman & Littlefield International, 2018.

Fazi, M. Beatrice. "Beyond human: deep learning, explainability and representation." *Theory, Culture & Society* 38, no. 7–8 (November 2020). https://doi.org/10.1177/0263276420966386.

Foucault, Michel. *Society Must Be Defended: Lectures at the Collège de France, 1975–1976*. Edited by Mauro Bertani and Alessandro Fontana. Translated by David Macey. London: Penguin Books, 2003.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning. Adaptive Computation and Machine Learning*. Cambridge: MIT Press, 2016.

Goodman, James, Sebastian Risi, and Simon Lucas. "AI and Wargaming." arXiv:2009.08922 [cs.AI] (2020).

Gregory, Derek. "From a view to a kill: Drones and late modern war." *Theory, Culture & Society* 28, no. 7–8 (2011): 188–215. https://doi.org/10.1177/0263276411423027.

Griffiths, Daniel. "Improbable's defence business awarded contract for a second year." *Improbable*, 12 November 2020. https://web.archive.org/web/20210304225936/https://www.improbable.io/blog/improbable-uk-strategic-command-sse/.

Hagendorff, Thilo. "The ethics of AI ethics: An evaluation of guidelines." *Minds and Machines* 30, no. 1 (2020): 99–120. https://doi.org/10.1007/s11023-020-09517-8.

Haggman, Andreas. "*Cyber Wargaming: Finding, Designing, and Playing Wargames for Cyber Security Education*." PhD Thesis, Royal Holloway, University of London,

2019. https://web.archive.org/web/20210703163737/https://pure.royalholloway.ac.uk/portal/files/33911603/2019haggmanaphd.pdf.

Hayles, N. Katherine. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago: University of Chicago Press, 1999.

Hayles, N Katherine. *Unthought: The Power of the Cognitive Nonconscious*. Chicago: University of Chicago Press, 2017.

Hoffman, Wyatt. "*AI and the Future of Cyber Competition*." Center for Security and Emerging Technology, 2021. https://web.archive.org/web/20210716152645/https://cset.georgetown.edu/wp-content/uploads/CSET-AI-and-the-Future-of-Cyber-Competition-4.pdf.

Holmqvist, Caroline. "Undoing war: War ontologies and the materiality of drone warfare." *Millennium* 41, no. 3 (2013): 535–552. https://doi.org/10.1177/0305829813483350.

Huelss, Hendrik. "Norms are what machines make of them: autonomous weapons systems and the normative implications of human-machine interactions." *International Political Sociology* 14, no. 2 (2020): 111–128. https://doi.org/10.1093/ips/olz023.

Hui, Yuk. *Recursivity and Contingency: Media Philosophy*. London: Rowman & Littlefield International, 2019.

Jensen, Benjamin M, Christopher Whyte, and Scott Cuomo. "Algorithms at war: The promise, peril, and limits of artificial intelligence." *International Studies Review* 22, no. 3 (2020): 526–550. https://doi.org/10.1093/isr/viz025.

Kania, Elsa B. "*Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power*." Center for a New American Security, 2017. https://web.archive.org/web/20210813153909/https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/Battlefield-Singularity-November-2017.pdf?mtime=20171129235805&focal=none.

Konda, Revanth, Hung Manh La, and Jun Zhang. "Decentralized function approximated Q-learning in multi-robot systems for predator avoidance." *IEEE Robotics and Automation Letters* 5, no. 4 (2020): 6342–6349. https://doi.org/10.1109/LRA.2020.3013920.

Lawson, Ewan, and Kubo Mačák. "*Avoiding Civilian Harm from Military Cyber Operations During Armed Conflicts*." Geneva, Switzerland: International Committee of the Red Cross, 2021. https://shop.icrc.org/download/ebook?sku=4539/002-ebook.

Le Cun, Yann. *Quand La Machine Apprend: La Révolution Des Neurones Artificiels et de l'apprentissage Profond*. Paris: Odile Jacob, 2019.

Lee, Dennis, Haoran Tang, Jeffrey Zhang, Huazhe Xu, Trevor Darrell, and Pieter Abbeel. "Modular architecture for starcraft ii with deep reinforcement learning." *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 14 (2018): 187–193.

Levinas, Emmanuel. *Totality and Infinity: An Essay on Exteriority*. Translated by Alphonso Lingis. The Hague: Martinus Nijhoff Publishers, 1979.

Liu, Ariel. "An introduction to machine learning." *DataSeries*, 16 September 2019. https://web.archive.org/web/20210914111118/https://medium.com/dataseries/a-brief-introduction-to-machine-learning-aeb55dae2288.

Lynch, Casey R., and Vincent J. Del Casino. "Smart spaces, information processing, and the question of intelligence." *Annals of the American Association of Geographers* 110, no. 2 (2020): 382–390. https://doi.org/10.1080/24694452.2019.1617103.

Mackenzie, Adrian. *Machine Learners: Archaeology of a Data Practice*. Cambridge: MIT Press, 2017.

Majumdar Roy Choudhury, Lipika, Alia Aoun, Dina Badaway, Luis Antonio de Alburquerque Bacardit, Yassine Marjane, and Adrian Wilkinson. "*Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011)*." New York: United Nations Security Council, 2021. https://web.archive.org/web/20210914000151/https://undocs.org/S/2021/229.

Ministry of Defence. "*Digital Strategy for Defence: Delivering the Digital Backbone and Unleashing the Power of Defence's Data*." London: UK Government, 2021. https://web.archive.org/web/20210917133641/https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/990114/20210421_-_MOD_Digital_Strategy_-_Update_-_Final.pdf.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves et al. "Human-level control through deep reinforcement learning." *Nature* 518, no. 7540 (2015): 529–533. https://doi.org/10.1038/nature14236.

Morley, Jessica, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander, and Luciano Floridi. "Ethics as a service: A pragmatic operationalisation of AI ethics." *Minds and Machines* 31 (June 2021): 239–256.

Nguyen, Thanh Thi, and Vijay Janapa Reddi. "Deep reinforcement learning for cyber security." arXiv:1906.05799 [cs.CR] (2019).

Parikka, Jussi. "Ethologies of software art: What can a digital body of code do?" In *Deleuze and Contemporary Art*, edited by Stephen Zepke and Simon O'Sullivan, 116–32. Edinburgh: Edinburgh University Press, 2010.

Parisi, Luciana. "Critical computation: Digital automata and general artificial thinking." *Theory, Culture & Society* 36, no. 2 (2019): 89–121. https://doi.org/10.1177/0263276418818889.

Ralston, William. "The untold story of a cyberattack, a hospital and a dying woman." *WIRED*, 11 November 2020. https://web.archive.org/web/20210916104511/https://www.wired.co.uk/article/ransomware-hospital-death-germany.

Schechter, Benjamin, Jacquelyn Schneider, and Rachael Shaffer. "Wargaming as a methodology: The international crisis wargame and experimental wargaming." *Simulation & Gaming* 52, no. 4 (2021): 513–526. https://doi.org/10.1177/1046878120987581.

Schrittwieser, Julian, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez. "Mastering atari, go, chess and shogi by planning with a learned model." *Nature* 588 (2020): 604–609. https://doi.org/10.1038/s41586-020-03051-4.

Schwartz, Jonathon, and Hanna Kurniawati. "Autonomous penetration testing using reinforcement learning." arXiv:1905.05965 [cs.CR] (2019).

Schwarz, Elke. "Autonomous weapons systems, artificial intelligence, and the problem of meaningful human control." *The Philosophical Journal of Conflict and Violence* 5, no. 1 (2021): 53–72. https://doi.org/10.22618/TP.PJCV.20215.1.139004.

Sharkey, Amanda. "Autonomous weapons systems, killer robots and human dignity." *Ethics and Information Technology* 21, no. 2 (2019): 75–87. https://doi.org/10.1007/s10676-018-9494-0.

Sheikh, Hassam Ullah, Mina Razghandi, and Ladislau Bölöni. "Learning distributed cooperative policies for security games via deep reinforcement learning." In

2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC). IEEE, 2019. https://doi.org/10.1109/COMPSAC.2019.00075.

Silver, David, and Demis Hassabis. "AlphaGo zero: Starting from scratch." *DeepMind*, 18 October 2017. https://web.archive.org/web/20210917133320/https://deepmind.com/blog/article/alphago-zero-starting-scratch.

Silver, David, Satinder Singh, Doina Precup, and Richard S. Sutton. "Reward is enough." *Artificial Intelligence* 299 (October 2021): 103535. https://doi.org/10.1016/j.artint.2021.103535.

Smeets, Max. "US cyber strategy of persistent engagement & defend forward: Implications for the alliance and intelligence collection." *Intelligence and National Security* 35 no. 3 (2020): 444–453. https://doi.org/10.1080/02684527.2020.1729316.

Smith, Roger. "The long history of gaming in military training." *Simulation & Gaming* 41, no. 1 (2010): 6–19. https://doi.org/10.1177/1046878109334330.

Stevens, Tim. *Cyber Security and the Politics of Time*. Cambridge: Cambridge University Press, 2016.

Stevens, Tim. "Knowledge in the grey zone: AI and cybersecurity." *Digital War* 1 (March 2020): 164–170.

Suchman, Lucy, Karolina Follis, and Jutta Weber. "Tracking and targeting: Sociotechnologies of (in) security" *Science, Technology, & Human Values* 42, no. 6 (2017): 983–1002. https://doi.org/10.1177/0162243917731524.

Sutton, Richard S, and Andrew G Barto. *Reinforcement Learning: An Introduction*. Cambridge: MIT Press, 2018.

Takaesu, Isao. "DeepExploit." *Github*, 2018. https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/DeepExploit.

Vinyals, Oriol, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi. "Grandmaster level in starcraft II using multi-agent reinforcement learning." *Nature* 575, no. 7782 (2019): 350–354. https://doi.org/10.1038/s41586-019-1724-z.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GPDR." *Harvard Journal of Law & Technology* 31, no. 2 (2017): 841–887.

Wang, Hanchao, Hongyao Tang, Jianye Hao, Xiaotian Hao, Yue Fu, and Yi Ma. "Large scale deep reinforcement learning in war-games." In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020. https://doi.org/10.1109/BIBM49941.2020.9313387.

3 Artificial intelligence in hybrid and information warfare

A double-edged sword

Wesley R. Moy and Kacper T. Gradon[1]

Generating and detecting misinformation

There are two complementary approaches in statistical classification for generating content using artificial intelligence (AI): the *Generative Adversarial Network* (GAN) approach and the *Discriminative* approach. The goal of the generative models is to produce synthetic data points that are randomly drawn from a wider distribution like the one responsible for generating real data points. For example, a generative model could generate fake photos of people that do not exist. Such photos, when analysed by a human or another machine learning model, would likely be classified as genuine human faces. On the other hand, discriminative models aim to distinguish between different kinds of data presentations. A discriminative model can be trained, for example, to distinguish a photo of a dog from a photo of a cat.

Generative models are more challenging to build than corresponding discriminative models. A generative model for images should be able to learn correlations like "boats are almost always surrounded by a large body of water" and "road bikes are rarely observed near a large pile of snow". Distinguishing cats from dogs seems to be a much easier task to do. Indeed, this intuitive discrepancy in difficulty may be explained. Discriminative models are supervised learning problems, in which the model is provided with labelled data to learn from, such as photos of cats and dogs divided into two corresponding piles. The challenge of building a generative model is that the model is expected to produce images that are normally considered as an input to traditional supervised learning algorithms, but it is not clear how such a process should be guided. A natural idea is to pass the output of a generational model to a discriminative model for evaluation. Introduced in 2014, this is what the GANs do. Two neural networks are trained at the same time, playing a zero-sum game in which one algorithm's gain is the other one's loss. The generative network generates candidates such as cats and dogs, and the discriminative network evaluates them. In other words, the generator tries to fool the discriminator, which at the same time learns to do its part – that is to recognize genuine images from the artificially constructed ones. An important property of GANs is that the networks can learn from each other

DOI: 10.4324/9781003284093-4

in an unsupervised way, that is, without providing any new data points such as labelled pictures of cats and dogs. As they progress through this iterative process, we can observe them improving their ability at a task.

GANs became very good in the computer vision domain but Natural Language Processing (NLP) problems are much more challenging, including generating and recognizing the textual dis- and misinformation. The main problem is that while text is a discrete object, as there is only a finite number of letters in the alphabet, the set of possible words is enormous. Specifically, it is difficult to back propagate the loss of signal from the discriminator back to the generator, that is, the discriminator cannot easily communicate with the generator to provide feedback on how well it did its job. The reason is that the *argmax* operation used by the generator to produce discrete symbols is non-differentiable. There are three ways to solve this issue, use reinforcement learning strategies, operate on continuous representations, or use the Gumbel-softmax operation.[2]

Some older and less effective approaches used previously, include Naive Bayes, Hidden Markov Models (HMMs), and plain Recurrent Neural Networks (RNNs). The current approach to generate synthetic, human-like text, however, is to use Large Language Models (LLMs) that are carefully pre-trained on enormous datasets. Generative Pre-trained Transformer 3 (GPT-3) created by OpenAI is a deep learning network with a capacity of 175 billion parameters, ten times more capacity than the second largest system, Microsoft's Turing NLG.[3] The main power of the OpenAI model comes from careful pre-training. To compare the scale of this process, the entire Wikipedia is part of the data used to train this model but amounts to just 3% of the hundreds of billions of words used for training.

GPT-3 is able to learn rapidly from a small set of training examples due to extensive pre-training and parameter size. An example of this would be to train GPT-3 to translate documents written in English to French by providing a few examples of English to French translations to the model and leaving the last example untranslated for GPT-3 to produce the translation. This task does not require tuning or weight updates and would not be possible without an extensive pre-training process. Its generic architecture can be applied to most tasks. It has been demonstrated that GPT-3 is able to learn 3-digit arithmetic and is also capable of coding in Python and other programming languages. Finally, GPT-3 excels at its primary task, generating text. GPT-3-generated texts are nearly indistinguishable from human-written texts. In one experiment, 80 native speakers were asked to identify which 200-word documents are generated by GPT-3 and which ones by humans. The subjects correctly distinguished GPT-3 generated texts approximately 52% of the time, which is not significantly higher than a random chance.

Humans interact with computers, cell phones, and other electronic devices in numerous ways. To attract customers and make their experience both natural and pleasant, developers of cell phone applications or computer software try to improve human-computer interaction (HCI) by including

natural interactions such as voice user interface (VUI) and enabling dialogs resembling conversation between people. Models such as GANs and LLMs are able to reduce the cognitive barriers between machines and people.

An early example of reducing cognitive barriers was designing interfaces for querying large databases using natural language. In such solutions, the user formulates a query for the type of products they are looking for in an online store, using common speech. The speech is interpreted into text and translated into formal query language. After the query is executed, a reverse process is performed. The result is first translated to a natural language sentence, which, in turn, is emitted as a voice response. A recent and advanced development of this kind is a product developed by Microsoft called Power Apps. It provides a low-code application development process in which you describe, in plain language, what you want to do and an application is quickly created.[4]

Following our customer of the online shop example, GANs can be used to dynamically create images of the offered products in the environment where they are naturally used. For example, instead of showing a traditionally stand-alone image of some piece of fashion, it can be adjusted to customer's needs and presented in various contexts of a person wearing it, including matches for gender, height, weight, and skin colour. It is important to highlight that although similar solutions are available without GANs they are limited because they are always custom-made and require significant involvement of human designers. As a result, they are labour intensive and expensive to create.

Another use of GANs and LLMs focuses on their core capability, the ability to generate synthetic objects that may be considered a sample from a distribution generating genuine instances. Such an application is described where GANs are used to generate synthetic orders that resemble genuine ones for e-commerce platforms such as Amazon, Alibaba, or Flipkart. Such a tool allows retailers to explore the space of plausible orders, and can provide insights into future customer orders, preferences, or price sensitivity.[5]

Since LLMs are trained on enormous and unlabelled datasets, they produce models that have bias against gender and marginalized groups that are present in the digested text. This raises valid criticism that should not be unexpected. This was noted in the original paper introducing GPT-3, which includes a detailed analysis of the biases observed including gender, race, and religion.[6] It is important to remember that the models are designed and expected to reflect the real world rather than how we would like the world to be. It is possible to alter the model to reduce bias, potentially making them less reflective of the real world and there are approaches to accomplish this.[7] GPT-3, and its earlier version GPT-2, exhibit bias against Muslims, marginalized groups, and other minorities.[8] This is true for other LLMs and pre-trained embeddings.[9] Occupational-based bias is also known to be present, associating women with occupations such as babysitting or nursing, and men with occupations such as construction workers and truck drivers.[10]

The ingested text will likely include inflammatory or hateful content. As a result, LLMs might generate not only biased text but also hateful and extremist content.[11] Unfortunately, it means that GPT-3 might be weaponized to create conspiracy, extremist, misleading or hateful content with little intervention by humans.[12] It has been suggested by several researchers that the datasets used to train these models should be carefully curated to reduce this possibility.[13]

Since LLMs try to "memorize" the training datasets, it might be possible to extract sensitive information from the model. For instance, with queries carefully designed for this purpose, one study extracted private information including phone numbers, full names, addresses, social media accounts, and more.[14] It may also be possible to generate sensitive national security information through a process recognized by intelligence agencies as *classification by compilation*.

Finally, let us mention that pre-training of LLMs often takes weeks of continuous computation done by hundreds of CPUs.[15] As a result, the cost to train the network can be estimated to be hundreds of thousands of dollars implying a sizable financial cost as well. This concern has been raised and some approaches for the optimization of the pre-training process have been proposed.[16]

There are several approaches to detect computer-generated text, most of them trying to detect style differences in which humans or generative models produce text. Such analyses work well on human-written text and can be used to detect the author of a specific document. The easiest approach is the *bag-of-words* classifier that computes how often given words are used in the document. Slightly more sophisticated approach is the *skip-gram* model, which tries to detect the context in which a given word was used. These techniques are only a start and more approaches are needed. As LLMs get better in mimicking patterns of humans and ensuring that no suspicious word sequence is present this might be used as evidence that the text was not written by a human.

Algorithms detecting generated text have quickly become ineffective, but research in this domain continues.[17] The state-of-the-art detectors GROVER[18] and RoBERTa[19] are approximately the same size as GPT-2, making it impossible for average researchers to use. It appears that to detect computer-generated text, access to computing power comparable to malicious actors generating the text is necessary. Moreover, it has been demonstrated that – at present – a simple statement inversion and other semantically neutral modifications can confuse the detector.[20] The problem of detecting generated text is extremely challenging but it may not be the issue that should be most concerning. A closely related question is to detect misinformation that aims to create a societal harm. This problem may be easier and there may be solutions.

The most natural approach to identifying misinformation is using Knowledge Graphs (KGs) to build a content-based detector. Unfortunately, KGs

are typically manually curated meaning that integrating new information typically requires several days. Misinformation, however, has the potential to cause significant damage in just a few hours. As a result, such solutions relying on fact checking sites such as Snopes or Politifact are not suitable for early detection algorithms.[21]

The U.S. National Security Commission on Artificial Intelligence noted that AI could be used to manipulate training data with the potential for compromising AI learning.[22] There are a few approaches for detecting misinformation. Unfortunately, it is difficult to judge which is the most effective as there is no standard benchmark dataset that might be used to make a fair comparison.[23] These approaches are often complex and are beyond the scope of this chapter. We will briefly outline them here.

The content-based approach uses content of the information rather than considering how that information is propagated. For example, a click bait-like format of a headline or sentiment polarity might be an effective discriminator. This approach has advantages as it can be used for early detection algorithms.[24] The social-response-based approach investigates how users react to a post. Responses such as "fake" or "untrue" indicate that users perceive misinformation. Emotional responses such as "unbelievable!" and political term such as "PC shit!"[25] also correlate with inaccurate information.[26]

A hybrid approach combines the social response of users with their characteristics. Each user is given a *suspiciousness score* based on their interaction with the system. The suspiciousness score of a piece of news is then simply a sum of the corresponding scores of the users interacting with the post. This approach was first used in Capture, Score, Integrate (CSI). Similar approaches were implemented in dEFEND and FNED.[27] The graph-based approach investigates how information spreads across a network. The underlying hypothesis is that inaccurate information spreads differently than accurate information and can be identified by investigating their pathways and the way users respond and interact with them. TraceMiner and FANG are two examples of successful algorithms using such techniques.[28] Finally, the multimodal approach incorporates information from multiple modalities to build better predictive models. The simplest example is to combine visual and textual information. This was shown to boost baseline accuracy by up to 7%.[29]

We continue to learn how social networks evolve and how users interact with each other, spreading misinformation and learning from each other. Tools are improving for understanding such networks. Since these processes are complex and interact with each other, often the only way to see the outcome of a given action is by performing extensive simulations. Through combining all of these techniques, we can try to detect malicious behaviour, understand the impact it has, or even strike back and counterbalance such actions.

Understanding the processes shaping social networks is an important task for social scientists trying to understand human behaviour but also for data scientists designing algorithms and tools working on those networks. Better

knowledge and dedicated tools may allow predicting which posts will be popular, which users have the power to change the shape of the network, which political view will likely become dominant, and, more importantly, how to detect if someone affects this processes and how to counterbalance their effects.

In the area of social network dynamics, there are several aspects that are relevant for understanding potential threats:

- How links between social agents are formed. What structures of links emerge naturally, and which structures are most likely to be engineered by malicious actors?
- How misinformation spreads in a network. Can the speed of its spread be predicted and influenced including identifying the social agents that are most likely to help spread misinformation.
- How access to information, sources, timing, dynamics, and form, influences the opinions of social agents. Specifically, how malicious agents can shape the listed factors to maximize the effectiveness of adversarial campaigns.

One of the first models of a dynamic network is the *preferential attachment model* introduced by Barabási and Albert,[30] who observed a power law degree sequence for a subgraph of the World Wide Web soon after the property was observed for the internet graph.[31] This property conforms to the *Pareto principle*, which posits that about 80% of effects result from 20% of causes. For example, the top 20% of earners in United States pay roughly 80–90% of Federal income taxes. The preferential attachment model explains why power-law degree distribution, displaying a high degree of asymmetry, occurs in many real-world complex networks such as the *rich-get-richer phenomenon*. The preferential attachment rule incorporated in the model makes new nodes to be more likely to connect to the more connected nodes than to the smaller nodes, creating a power-law degree distribution.

There are many other important models explaining various properties such as increasing edge density (densification) and decreasing diameter or surprisingly large clustering coefficient and short average path lengths, the properties explained by the Watts-Strogatz model.[32] Below we concentrate on a few aspects that need identification.

For detecting malicious activity, the key aspects of social network analysis are:

- Certain local structures of connections between nodes in the social graph that can be considered as suspicious, often referred to as motifs.
- Overly regular sub-graphs of a social network. Empirical analyses show that social ties are characterized by the presence of certain levels of noise.
- Anomalous nodes with vastly different characteristics in comparison to what would be predicted for them.

Spreading misinformation

Social networks create an information platform in which human and software-assisted accounts (bots) try to take advantage of the system for various reasons including triggering collective attention,[33] gaining status,[34] monetizing public attention,[35] spreading disinformation, or seeding discord.[36]

A large number of Twitter accounts are bots that are responsible for much of the disinformation shared. Recent studies showed that bots play an important role in the initial stage of diffusion by amplifying low-credibility content, but they cannot distinguish between true and false information. Humans, however, are more likely to spread false information.[37] It is possible to characterize the behaviour of individuals who use bots to enhance their online visibility and influence. Such accounts, bot-assisted humans or human-assisted bots, are often referred to as *cyborgs* or *augmented humans*.[38] Opportunities for spreading (mis-, dis-)information and banning strategies depend highly on whether a social platform is moderated or not.[39] Finally, various countries exhibit different levels of infodemic risk, adding yet another level of complexity.[40]

Understanding of social networks has advanced significantly, but the underlying mechanisms for spreading false information are still not well understood. Recent results suggest that spreading mechanisms might have indistinguishable population-level dynamics.[41] We are flooded with information at a level impossible to consume. Human attention is limited and individual reaction to information is a complex interplay between individual interests and social interaction. Collective attention is typically characterized by a quickly growing focus on a specific topic, such as presidential elections or vaccination, until a well-identified peak of attention is reached. This is followed by the second phase in which a slow decay of interest is observed.[42]

Initial research neglected the effects of the underlying social structure but it is clear that underlying network structure plays an important role in the process.[43] De Domenico and Altmann combine two simple mechanisms to explain the dynamic of collective attention: a preferential attachment process shaping the network topology and a preferential attention process responsible for individual's attention bias towards specific users of the network.[44]

Social learning refers broadly to the processes by which a person's social environment shapes how a person behaves and thinks. In the context of social networks, a user may adopt the cognitions or behaviours from those they interact with directly. Concurrently, learning also shapes the social environment, since individuals also exercise control over their social environment and potentially select network partners as a function of individual attributes including behaviours and cognitions.[45] As a result, social learning is a complex process but important in understanding the spread of misinformation. To illustrate its complexity, we focus on two aspects: segregation and polarization.

Various models explain why and how segregation occurs beginning with the classic theory of residential segregation.[46] Segregation theory also applies

to the structure of networks; segregated networks always emerge even if the users are assumed to have only a small aversion from being connected to others who are dissimilar to themselves and yet no actor strictly prefers a segregated network.[47] The power of aversion is often amplified by homophily, a tendency for people to have ties with similar others.[48] For example, Bener et al. developed and tested empirical models of how social networks evolve over time; particularly, how people in a social network choose links on the basis of their own attributes, and how individual attributes are in turn shaped by network structure.[49]

An extreme situation, polarization, occurs when a group is divided into two opposing sub-groups having conflicting and contrasting positions, goals, and points of view with few individuals remaining neutral. Polarization typically occurs in politics but several other issues often experience it including climate change, gun control, same-sex marriage, and abortion. The presence of polarization changes the dynamics of a network. Indeed, for example it is known that Twitter users are unlikely to be exposed to cross-ideological content from the clusters of users they followed, as these groups were usually politically homogeneous.[50] Antonakaki, Fragopoulou and Ioannidis identified this in a recent survey of Twitter research.[51]

Hypergraphs

Most network science concentrates on building tools for mining complex networks represented as a graph. Facebook, for example, might be represented as a set of nodes associated with users and a set of edges between nodes associated with friendship relationship between the users. Most of the social media platforms, especially forums, are more complex and require more sophisticated representations to properly represent them. Interactions between users commenting on a given post may be represented as a tree rooted at the node associated with the user who posted the initial post and internal nodes in the tree associated with users who comment on someone else's comment.

Current technologies and tools are not capable of dealing with such complex structures but there is movement to the next level introducing tools that deal with hypergraphs. In this simpler structure, nodes form hyperedges that may consist of any number of members, not only two as in the case of graphs. A hyperedge may include all users that interacted with a specific post, or "liked" a particular photo on Instagram. There has been a recent surge of interest in higher-order methods, especially in the context of hypergraph clustering.[52] Battiston et al. provided a recent study on the higher-order architecture of real complex systems,[53] and for a hypergraph neural networks framework (HGNN) for data representation learning.[54]

Another recent effort has been to represent each node of a network as a low dimensional feature vector. There are a number of reasons for using this approach. On the one hand, it decreases the dimension of the problem making it easier to handle and analyse. Moreover, even though dimensionality

reduction means losing some information, it is actually a useful technique. Indeed, well-trained algorithms aim to learn the most important information about the network but ignore "noise". As a result, "compressed" representation is often more useful and informative than the complete picture (an analogue situation would be as follows: reading a relatively short summary of the two-hours long meeting might be more useful than listening to the original recording). As a result, over 100 node-embedding algorithms have been proposed recently in the literature. The techniques and possible approaches to construct the desired embedding can be broadly divided into the following families: linear algebra algorithms, random walk-based methods, and deep learning methods.[55] Selecting an appropriate embedding for a given network and a given task at hand typically require ad-hoc experiments and tests performed by domain experts. However, a general framework was recently proposed that provides a tool for an unsupervised graph embedding comparison, which should make the selection process easier, and the outcome of better quality.[56]

The network dynamics and information spread in them is complex. We have highlighted a few aspects, but this is only a starting point. This complexity is driven by several factors. We are faced with a very large number of agents in the system. Each has its own characteristics making them distinct. In the literature this feature is called agent heterogeneity.[57] The behavioural rules of agents are typically complex, depend on their characteristics, and evolve over time. The description of the microstructure of the social network is quite extensive. In most cases, we are interested in some macro consequences of this microstructure such as the speed and reach of malicious information. Experience in modelling of enormous systems show that even small changes in microstructure can lead to complex and non-obvious changes in macro results. In the literature, such consequences are called emergent, as they are frequently impossible to deduce given only the microstructure assumptions.[58]

A typical approach to address these challenges is to construct a *digital twin*, a synthetic model of a real system.[59] The representation of the real social network in such an approach is usually done in 1:1 scale. The calibration of the microstructure of the system is made based on available individual and aggregated data although it is sometimes impossible to collect all relevant characteristics of the system on an individual agent level. Such a representation is done using software implementation.[60] Given the scale of the simulations the crucial requirements are high-performance tools used to create such models.

Virtual simulators of the social network interactions allow for performing flexible what-if analysis that provides verification of how the system would react to different stimuli. It is important to stress here that very often the modelled systems are highly volatile and the response of the system is hard to predict. Therefore, the simulations usually are run multiple times creating a distribution of potential outcomes rather than a single point estimate.[61] Such approaches are commonly used to find optimal seeding of social media

campaigns and find the agents in the social network who should be used to initialize spreading of information to ensure its maximum reach and speed.[62]

The simulation approach is often coupled with *metamodeling*, also called *surrogate modelling*.[63] The metamodel of the simulation provides a simplified view and typically uses supervised learning techniques and *Explainable AI* (XAI) tools.[64] The objective of creating of a metamodel is twofold. First, analysts working with a complex simulation model are usually interested in which components determine the direction and impact of the outcomes. The goal of the metamodel is to provide a simplified view of the complex system in which accuracy is traded for easier interpretation and greater understanding. Such qualitative insights are often invaluable for understanding the grand principles that control and influence social networks can apply. The second use of the metamodels is related to speed with which they can produce predictions. Often the original simulation model can take hours or days to produce outcomes. In contrast, metamodels can produce results within seconds, albeit with less precision, a trade-off between speed and accuracy. Such a trade-off is often desirable in applications where alternative courses of action need to be evaluated and action taken quickly.[65]

Conflict among the major competitors

COVID-19 has been used to advance nation-state political agendas and support their efforts in competition and conflict. The health emergency has enabled campaigns bearing the characteristics of hybrid warfare with the intelligence services of both the European Union and United States recognizing the intensified activity of foreign actors.[66] The use of asymmetric methods or hybrid warfare techniques and procedures should be viewed as activities along a spectrum of conflict short of kinetic warfare. These methods attack both the target state's population as well as attack the decision-making process of the target state.[67] The concept of hybrid warfare has been used to refer to the combination of military tactics such as conventional warfare with irregular warfare and cyber warfare, as well as the employment of other subversive instruments and tactics, to achieve two goals: first, to avoid responsibility and retribution, and second to weaken and destabilize the enemy without perceived involvement.[68] Hybrid warfare influence campaigns include propaganda, malinformation, and disinformation to influence the thoughts and beliefs of the target population.[69] A defining element of hybrid warfare is targeting of the population alongside traditional political, economic, and information aspects of conflict.[70] The pandemic crisis has been used to spread misinformation and disinformation to further political objectives and expand national influence.[71] Early in the crisis, both Russia and China have used the pandemic to further their political goals such as sending medical advisors and supplies to Italy.[72] The Pew Research Center found that prominent officials and news media outlets in Russia, China, Iran, and the United States became super-spreaders of disinformation about COVID-19.[73] Russian and

Chinese media efforts consistently seek to undermine confidence in Western developed vaccines including reporting false information including sensationalizing safety concerns and promoting the narrative that their vaccines are superior to those produced in the West.[74] A significant amount of misinformation related to the pandemic has come from credible sources increasing the problem of distinguishing accurate from inaccurate information.[75] AI includes powerful tools to both generate misinformation and detect its use. As Kasapoglu and Kirdemir note, AI could cause drastic changes in hybrid warfare, which is a major concern for NATO.[76] States and non-state actors can use cyberspace to influence large groups of civilians and opposing forces. From reconnaissance activities and the profiling of target audiences to the weaponization of distorted or fake information and psychological operations, AI broadens the potential of information operations. The problem being considered is how to combat misinformation in cyber space using AI to identify sources, discover how misinformation travels, and counter its adverse effects. Concurrently, AI can be weaponized in information operations making it both a weapon and a threat.

Infodemics is a term first used to describe the information overflow during the 2003 Severe Acute Respiratory Syndrome (SARS) epidemic.[77] In the beginning of the COVID-19 pandemic, the World Health Organization (WHO) noted that the health crisis was accompanied and amplified by the astonishing data overload and unprecedented information chaos, developing at the extraordinary pace, and began to use the term of Infodemics to address it.[78] The WHO also called to unify the vocabulary related to Infodemics, endorsing the creation of a unified lexicon allowing for the clear understanding and use of terminology among the community of experts.[79] Such approach was supported by the subject matter experts[80] and the following vocabulary, adapted from Wardle and Derakhshan,[81] has been in use among the infodemic management community since.

Disinformation is false information that is deliberately created or disseminated with the express purpose to cause harm; fabricated content and the malicious intent behind it are the identifying properties of disinformation. The term *misinformation* is frequently and erroneously used as a synonym of disinformation, but, although the disseminated information is false, it is not intended to cause harm, as the persons sharing it believe that it is true and accurate and attempting to be helpful. Finally, *malinformation* is genuine information that is broadcasted with a deliberate intent to cause harm, by revealing data that may hurt the reputation of a person or an institution.[82] It is important to highlight, that infodemiology experts reject the term *fake news*, which was coined to describe the use of disinformation and misinformation in news reporting.[83] The term has been used by political actors to discredit news reporting and reported facts they dislike.[84]

The COVID-19 pandemic has highlighted conflict in cyberspace among three major competitors: Russia, China, and the United States. Russia and its predecessor, the Soviet Union, have had a longstanding adversarial

relationship with the United States and its Western allies. China, in its quest to become the predominant state power, has utilized its cyber capabilities to advance all of its instruments of national power. In the United States, many aspects of the pandemic have become politicized, often used as weapons between the major political parties over measures to contain the virus and vaccinations against it. In the process, the United States has become a major source of misinformation that affects much of the world. Each of the major powers makes significant contributions to conflict in cyber space in its own way and for different reasons.

Russia has maintained a prolonged effort to undermine the legitimacy of Western democracy and democratic societies.[85] Putin perceives that the greatest threat to Russia is Western democracy and attacks the West to shift attention away from weaknesses including corruption and the economy.[86] The 2012 plan to increase economic growth was poorly conceived and lacked follow through.[87] The gross domestic product (GDP) for 2020 is estimated to be about $1.5 trillion, 11th in the world; however, per capita GDP was just $11,654 or 64th of all countries.[88] Russian economic and political power is concentrated in the hands of a small number of individuals made up of former intelligence officers and oligarchs.[89]

Russia regularly conducts covert action including spreading misinformation to exploit societal divisions and undermine democratic legitimacy.[90] Russian use of a health crisis to further its political agenda is not new. During the 1980s, Soviet intelligence disseminated disinformation that HIV/AIDS virus was developed by the United States as a biological weapon.[91] Russia has echoed Chinese claims that COVID-19 was developed and spread by the United States as a weapon against China.[92] Russian trolls have amplified disinformation to exacerbate the American debate on vaccines and undermine American politics.[93]

Russian Federation disinformation is a descendant of USSR's political warfare theories and practices.[94] Soviet KGB operations included subversion, media manipulation, propaganda, political repression, political assassination, the establishment of opposition parties and criminal organizations.[95] They are similar to the current practices of the Russian Federation's Federal Security Service (FSB) using political assassination, political tyranny at home, and sponsorship of paramilitary groups in independent nations. Their methods also include the use of soft power such as worldwide disinformation campaigns, foreign political meddling, and the establishment of networks of influence abroad.[96]

Putin's strategies use a combination of repression at home, disinformation and destabilization campaigns in the West, and military interventions in Ukraine and Syria. All result from the simultaneous strength and insecurity of the Russian leader, who struggles to manage a corrupt, kleptocratic, mafia style state.[97] Perhaps the Kremlin's policies are not a part of a grand strategy, but an opportunistic foreign policy. Putin probes for Western weakness, irresolution, and indecision, and then, if there is no resistance, he intervenes

to extend Russia's reach by absorbing, either physically, or by exerting influence, more territory.[98] Russia has begun to devote extraordinary resources to a political and information war designed to undermine and eventually destroy Western democracies and institutions.[99] Moscow is pursuing investments in high-impact, low-cost asymmetric warfare to correct the imbalance between Russia and the West in the conventional domain.[100]

China aspires to be the leading economic, military, and political player in the world, seeking to surpass the United States and its Western allies in all aspects of national power by 2035.[101] The U.S. Department of State has identified China as the central threat that undermines the stability of the world, identifying issues including predatory economic practices, disregard for human rights, and undermining global norms and values.[102] Made in China 2025 is an ambitious plan to upgrade Chinese industries, especially in advanced fields including robotics, AI, and quantum computing. The aggressive nature of the plan may motivate the use of questionable trade practices including cyber theft.[103] To reach its 2035 goals, China must continue to raise its industrial productivity through capital inputs, human resources improvements, and acquiring advanced technologies.[104] China's ability to reach this aggressive goal is in question.[105] The GDP for 2020 is estimated to be about $15 trillion, 2nd in the world; however, per capita was just $11,819 or 61st of all countries.[106] This may lead to even more aggressive efforts to acquire advanced technologies.

China uses a network of related media organizations to propagate misinformation online and harass opponents at home and abroad. It produces and echoes misinformation and disinformation on issues including election voter fraud, COVID-19, and QAnon theories.[107] China is modernizing all aspects of the People's Liberation Army from strategic nuclear forces and cyber capabilities to conventional capabilities of all types.[108] China continues to conduct espionage against the United States on a massive scale. They have been successful in stealing designs for advanced weapons systems including the F-22 and F-35 fighter jets and C-17 transport aircraft.[109] Recently, Chinese fighter jets have surpassed Russian aircraft performance, in part, due to their willingness to conduct industrial espionage and reverse engineer advanced technologies.[110] Chinese pharmaceutical companies are behind their Western competitors on mRNA technology, used to rapidly produce COVID-19 vaccines.[111] Two Chinese government-linked hackers attempted in 2020 to steal data related to mRNA technology from pharmaceutical firm Moderna.[112]

In the United States' politically polarized society, news outlets, fringe media and conspiracy sites – some with significant global reach – mislead their audiences with false narratives with significant negative outcomes.[113] According to a U.S. national survey during the early days of the COVID-19 pandemic, the respondents' identified political party, correlated with specific beliefs about protection from infection. Conservative media use, however, correlated with belief in conspiracy theories.[114] Additionally, a 2021 survey by the U.S. National Academy of Sciences found that 75% of persons

overestimated their ability to identify misinformation and the most overconfident were poorest at it and most likely to spread the information.[115]

The U.S. information chaos related to the COVID-19 pandemic was amplified by propagation of conspiracy theories.[116] Theories ranged from linking the disease with the 5G communications through QAnon community claims that the virus was created the government officials and other figures that secretly run the world. Such theories generated political and social effects, and increased anxiety, uncertainty, distrust towards authorities and fear.[117] Conspiracy theories in the QAnon community have posited that the coronavirus might not be real and that if it was, it had been created the government officials and other figures who secretly run the world.

COVID-19 misinformation in the United States came from both official and unofficial sources with devastating effects there and potentially worldwide. Evanega et al. found that about one third of English language COVID-19 misinformation were news media mentions of statements by U.S. President Donald Trump, of which just 16% was fact checked.[118] The rest of the media content was provided without question or confirmation. Brennen et al. observed that prominent public figures including politicians and celebrities accounted for 20% of pandemic misinformation, however, these individuals accounted for 69% of the misinformation echoed on social media.[119] The U.S.-based Center for Countering Digital Hate found that just 12 people in the U.S. were responsible for about two thirds of misinformation supporting vaccine hoaxes and vaccine hesitancy. These persons include anti-vaccine activists, doctors, and alternative health promoters whose organization generate up to $1 billion annual from these activities.[120] Belief in misinformation related to COVID-19 is associated with vaccine hesitancy and noncompliance with health guidance.[121]

In 2018, Google allow a contract with the U.S. Department of Defense to lapse because of employee opposition.[122] Operation Maven was intended to use AI with drone surveillance video to identify targets and support operations against them.[123] In March 2021, the U.S. National Security Commission on Artificial Intelligence identified potential uses of AI in conflict including AI enabled information operations. These threats include autonomous disinformation campaigns and computational propaganda including deepfakes.[124] There has been, however, some backlash against the use of AI in warfare.

Conclusions

Considered together, modelling can help to understand how links between agents are formed, how information is disseminated in a network, and how this information, combined with other factors, can influence opinions and actions of agents active in social networks. Together they have significant potential to blunt the effects of misinformation related to the pandemic as well as other types of misinformation campaigns such as anti-vaccination efforts.

Malicious agents can use the tools, usually with the goal of inciting some behaviour by people engaged in social network. Therefore, they use the modelling of the system to plan their actions to maximize impact. Typical actions in this area would involve injecting bots into a network, hijacking neutral nodes, or interfering in the communication between nodes. All the actions are typically aimed at distorting the structure and information flow within a network. The strategies used here are usually staged, during an initial phase the malicious nodes would act "normally" and, only after gaining credibility, start their detrimental activities.

These same tools can be used to fight malicious agents. The tools and techniques discussed can be used to detect anomalous activities in networks or network structures. When adversarial information is identified using either human analysis or analytical tools, the simulation and social network analysis techniques can be employed to identify sources of malicious actions within the network. Identification of the sources of misinformation may blunt the effectiveness of their messages and erode the perception of the veracity of what they disseminate.

Use of these AI tools has the potential to identify sources of misinformation, study how the misinformation moves through networks, and ultimately blunt the effects of the misinformation. We recommend testing these methodologies against the misinformation surrounding COVID-19 and the vaccinations against the virus. Several the sources of misinformation on both are known, allowing for validation of the AI techniques and training AI tools while supporting combating the misinformation and disinformation related to the pandemic. This effort will likely uncover additional false narratives as well as identify additional malicious actors. It will also increase acceptance of AI methodologies by nation-states, civil society, and individual consumers of information.[125]

Notes

1 The views expressed in this chapter are those of the authors and do not reflect the official policy or position of the Department of Homeland Security (DHS) or the U.S. Government. DHS cannot attest to the substantive or technical accuracy of the information. The final version of the chapter was submitted in February 2022 and represents the technical state of the art on that date.
2 R.J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning* 8, no. 3–4 (1992); S. Rajeswar et al., "Adversarial generation of natural language," arXiv:1705.10929 [cs.CL] (2017); M.J. Kusner and J.M. Hernández-Lobato, "Gans for sequences of discrete elements with the gumbel-softmax distribution," arXiv:1611.04051 [stat.ML] (2016).
3 T.B. Brown et al., "Language models are few-shot learners," arXiv:2005.14165 [cs.CL] (2020).
4 R. Cunningham, "Introducing power apps ideas: AI-powered assistance now helps anyone create apps using natural language," *Microsoft Power Apps* blog (25 May 2021).
5 A. Kumar, A. Biswas, and S. Sanyal, "eCommerceGAN: A generative adversarial network for e-commerce," arXiv:1801.03244 [cs.LG] (2018).

6 Brown et al., "Language models."
7 Brown et al., "Language models;" M. Kaneko and D. Bollegala, "Gender-preserving debiasing for pre-trained word embeddings," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics, 2019); T. Bolukbasi et al., "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," arXiv:1607.06520 [cs.CL] (2016); T. Manzini et al., "Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings," arXiv:1904.04047 [cs.CL] (2019).
8 A.M. Abid, M. Farooqi, and J. Zou, "Persistent anti-Muslim bias in large language models," arXiv:2101.05783 [cs.CL] (2021); P. Schramowski et al., "Language models have a moral dimension," preprint (2021); E.M. Bender et al., "On the dangers of stochastic parrots: Can language models be too big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2021).
9 W. Guo and A. Caliskan, "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases," arXiv:2006.03955 [cs.CY] (2020).
10 H. Kirk et al., "How true is GPT-2? An empirical analysis of intersectional occupational biases" (2021).
11 Bender et al., "Dangers of stochastic parrots."
12 K. McGuffie and A. Newhouse, "The radicalization risks of GPT-3 and advanced neural language models," arXiv:2009.06807 [cs.CY] (2020).
13 A. Tamkin et al., "Understanding the capabilities, limitations, and societal impact of large language models," arXiv:2102.02503 [cs.CL] (2021); Bender et al., "Dangers of stochastic parrots."
14 N. Carlini et al., "Extracting training data from large language models," arXiv:2012.07805 [cs.CR] (2020).
15 Brown et al., "Language models."
16 Bender et al., "Dangers of stochastic parrots;" Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692 [cs.CL] (2019).
17 G. Jawahar, M. Abdul-Mageed, and L.V.S. Lakshmanan, "Automatic detection of machine generated text: A critical survey," arXiv:2011.01314 [cs.CL] (2020).
18 R. Zellers et al., "Defending against neural fake news," arXiv:1905.12616 [cs.CL] (2019).
19 I. Solaiman et al., "Release strategies and the social impacts of language models," arXiv:1908.09203 [cs.CL] (2019).
20 T. Schuster et al., "The limitations of stylometry for detecting machine-generated fake news," *Computational Linguistics* 46, no. 2 (2020).
21 X. Zhou et al., "Fake news early detection: A theory-driven model," *Digital Threats: Research and Practice* 1, no. 2 (2020); Y. Liu and Y.-F.B. Wu, "Fned: A deep network for fake news early detection on social media," *ACM Transactions on Information Systems (TOIS)* 38, no. 3 (2020).
22 U.S. National Security Commission on Artificial Intelligence, *Final Report*, 2021.
23 B. Guo et al., "The future of false information detection on social media: New perspectives and trends," *ACM Computing Surveys (CSUR)* 53, no. 4 (2020).
24 Zhou et al., "Fake news early detection;" S. Kwon, M. Cha, and K. Jung, "Rumor detection over varying time windows," *PloS One* 12, no. 1 (2017); K. Shu et al., "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter* 19, no. 1 (2017).
25 PC is an abbreviation of "political correctness."

26 J. Ma et al., "Detecting rumors from microblogs with recurrent neural networks," *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ed. Subbarao Kambhampati (AAAI Press / International Joint Conferences on Artificial Intelligence, 2016): 3818; T. Chen et al., "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in *Pacific-Asia conference on knowledge-discovery and data mining*, eds. M. Ganji, L. Rashidi, B. Fung and C. Wang (Springer, 2018): 40–52.
27 N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *Proceedings of the 2017 ACM Conference on Information and Knowledge Management* (2017); Shu et al., "Fake news detection;" Liu and Wu, "Fned: A deep network."
28 L. Wu and H. Liu, "Tracing fake-news footprints: Characterizing social media messages by how they propagate," in *Proceedings of the eleventh ACM International Conference on Web Search and Data Mining* (New York: Association for Computing Machinery, 2018); V.-H. Nguyen et al., "FANG: Leveraging social context for fake news detection using graph representation," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (2020).
29 Z. Jin et al., "Novel visual and statistical image features for microblogs news verification," *IEEE Transactions on Multimedia* 19, no. 3 (2016).
30 A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science* 286, no. 5439 (1999).
31 M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *The Structure and Dynamics of Networks* (Princeton, NJ: Princeton University Press, 2011).
32 J. Leskovec, J., Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, no. 1 (2007); D.J. Watts and S.H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature* 393, no. 6684 (1998).
33 J. Lehmann et al., "Dynamical classes of collective attention in Twitter," in *Proceedings of the 21st International Conference on World Wide Web* (New York: Association for Computing Machinery, 2012); M. De Domenico and E.G. Altmann, "Unraveling the origin of social bursts in collective attention," *Scientific Reports* 10, no. 1 (2020).
34 M. Cha et al., "Measuring user influence in Twitter: The million follower fallacy," in *Proceedings of the International AAAI Conference on Web and Social Media* (2010); M. Stella, M. Cristoforetti, and M. De Domenico, "Influence of augmented humans in online interactions during voting events," *PloS One* 14, no. 5 (2019).
35 D. Carter, "Hustle and Brand: The Sociotechnical Shaping of Influence," *Social Media and Society* 2, no. 3 (2016).
36 C.A. Bail et al., "Assessing the Russian internet research agency's impact on the political attitudes and behaviors of American Twitter users in late 2017," *Proceedings of the National Academy of Sciences* 117, no. 1 (2020); D. Freelon et al., "Black trolls matter: Racial and ideological asymmetries in social media disinformation," *Social Science Computer Review* (2020); S.C. Woolley and P.N. Howard (eds), *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media* (New York: Oxford University Press, 2018).
37 S. González-Bailón and M. De Domenico, "Bots are less central than verified accounts during contentious political events," *Proceedings of the National Academy of Sciences* 118, no.11 (2021).
38 Stella, Cristoforetti and De Domenico, "Influence of augmented humans."
39 O. Artime et al., "Effectiveness of dismantling strategies on moderated vs. unmoderated online social platforms," *Scientific Reports* 10, no. 1 (2020).

40 R. Gallotti et al., "Assessing the risks of 'infodemics' in response to COVID-19 epidemics," *Nature Human Behaviour* 4, no. 12 (2020).
41 L. Hébert-Dufresne, S.V. Scarpino, and J.-G. Young, "Macroscopic patterns of interacting contagions are indistinguishable from social reinforcement," *Nature Physics* 16, no. 4 (2020).
42 Lehmann et al., "Dynamical classes;" De Domenico and Altmann, "Unraveling the origin."
43 J.P. Gleeson et al., "Effects of network structure, competition, and memory time on social spreading phenomena," *Physical Review* X6, no. 2 (2016).
44 De Domenico and Altmann, "Unraveling the origin."
45 S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proceedings of the National Academy of Sciences* 106, no. 51 (2009).
46 T. Schelling, "Models of segregation," *The American Economic Review* 59, no. 2 (1969).
47 A.D. Henry, P. Prałat, and C.-Q. Zhang, "Emergence of segregation in evolving social networks," *Proceedings of the National Academy of Sciences* 108, no. 21 (2011).
48 D. Byrne, "An overview (and underview) of research and theory within the attraction paradigm," *Journal of Social and Personal Relationships* 14, no. 3 (1997).
49 A.B. Bener et al., "Empirical models of social learning in a large evolving network," *PloS One* 11, no. 10 (2016).
50 I. Himelboim, S. McCreery, and M. Smith, "Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter," *Journal of Computer-Mediated Communication* 18, no. 2 (2013).
51 D. Antonakaki, P. Fragopoulou, and S. Ioannidis, "A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks," *Expert Systems with Applications* 164 (2021).
52 B. Kamínski et al., "Clustering via hypergraph modularity," *PloS One* 14, no. 11 (2019).
53 F. Battiston et al., "Networks beyond pairwise interactions: structure and dynamics," *Physics Reports* 874 (2020).
54 Y. Feng et al., "Hypergraph neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence* (2019).
55 W.L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," arXiv:1709.05584 [cs.SI] (2017); D. Zhang et al., "Network representation learning: A survey," *IEEE Transactions on Big Data* 6, no. 1 (2018); B. Kamínski, P. Prałat, and F. Théberge, *Mining Complex Networks* (New York: CRC Press, 2021).
56 B. Kamínski, P. Prałat, and F. Théberge, "An unsupervised framework for comparing graph embeddings," *Journal of Complex Networks* 8, no. 5 (2020).
57 M. Gallegati and M.G. Richiardi, "Agent based models in economics and complexity," in *Encyclopedia of Complexity and Systems Science* (New York: Springer, 2009).
58 E. Bonabeau, "Agent-based modelling: Methods and techniques for simulating human systems," *Proceedings of the National Academy of Sciences* 99, suppl. 3 (May 2002).
59 P. Raj and E. Preetha, *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases* (Elsevier, 2020).
60 N. Gilbert, *Agent-Based Models* (SAGE, 2008).
61 B. Kamínski, "A method for the updating of stochastic kriging metamodels," *European Journal of Operational Research* 247, no. 3 (2015).
62 M. Will et al., "Combining social network analysis and agent-based modelling to explore dynamics of human interaction: A review," *Socio-Environmental Systems Modelling* 2 (2020).

63 R. Barton, "Tutorial: Simulation metamodeling," *2015 Winter Simulation Conference (WSC)* (2015).
64 D. Rothman, *Hands-On Explainable AI (XAI) with Python* (O'Reilly, 2020).
65 B. Kamínski, "Interval metamodels for the analysis of simulation input-output relations," *Simulation Modelling Practice and Theory* 54 (2015).
66 K. Gradon et al., "Countering misinformation: A multidisciplinary approach," *Big Data & Society Special Issue on Studying Infodemic at Scale* 89, no. 1 (May 2021).
67 T.A. Schnaufer, "Redefining hybrid warfare: Russia's nonlinear war against the west," *Journal of Strategic Security* 10, no. 1 (2017).
68 A.C. Apetroe, "Hybrid warfare: from war during peace to neo-imperialist ambitions. The case of Russia," *On-line Journal Modelling the New Europe* 21, no. 1 (2016).
69 C. Atkinson, "Hybrid warfare and societal resilience: Implications for democratic governance," *Information & Security* 39, no.1 (2018).
70 R. Munteanu, "Hybrid warfare: The new form of conflict at the beginning of the century," *Strategic Impact* 56, no. 3 (2015): 19.
71 D. Broniatowski et al., "The covid-19 social media infodemic reflects uncertainty and state-sponsored propaganda" (2020).
72 E. Braw, "Beware of bad samaritans," *Foreign Policy*, 30 March 2020.
73 E. Kinetz, "COVID conspiracy shows vast reach of Chinese disinformation," *The Associated Press*, 15 February 2021.
74 R. Emmott, "Russia, China sow disinformation to undermine trust in western vaccines: EU," *Reuters*, 28 April 2021.
75 Broniatowski et al., "The covid-19 social media infodemic."
76 C. Kasapoglu and B. Kirdemir, "Artificial intelligence and the future of conflict," in *New Perspectives on Shared Security: NATO's Next 70 Years* (Brussels: Carnegie Europe, 2019).
77 D.J. Rothkopf, "When the buzz bites back," *The Washington Post*, 11 May 2003, B01.
78 World Health Organization (WHO), *Coronavirus Disease 2019 (COVID-19). Situation Report* 45, (2020).
79 World Health Organization (WHO), *1st WHO Infodemiology Conference: How Infodemics Affect the World & How They can be Managed*, Conference booklet (2020).
80 L. Turcilo and M. Obrenovic, "Misinformation, disinformation, malinformation: causes, trends, and their influence on democracy," *Heinrich Böll Foundation Companion to Democracy* 3 (August 2020).
81 C. Wardle and H. Derakhshan, "Information disorder: Towards an interdisciplinary framework for research and policy-making," *Council of Europe Report*, 27 September 2017.
82 Ibid.
83 A. Kucharski, *The Rules of Contagion: Why Things Spread – and Why They Stop* (New York Basic Books, 2020).
84 E. Carmi et al., "Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation," *Internet Policy Review* 9, no. 2 (2020).
85 W.R. Moy and K. Gradon, "COVID-19 effects and Russian disinformation," *Homeland Security Affairs* 16 (2020).
86 J. Biden and M. Carpenter, "How to stand up to the Kremlin: defending democracy against its enemies," *Foreign Affairs* 87, no. 1 (January/February 2018).
87 T. Frye, "Russia's weak strongman: The perilous bargains that keep Putin in power," *Foreign Affairs* 100, no. 3 (2021).
88 Statistics Times, *World GDP Ranking 2021* (2021).
89 Emmott, "Russia, China sow disinformation."
90 Ibid.
91 T. Boghardt, "Soviet bloc intelligence and its AIDS disinformation campaign," *Studies in Intelligence* 53, no. 4 (December 2009).

92 Moy and Gradon, "COVID-19 effects."
93 Broniatowski et al., "The covid-19 social media infodemic."
94 J.V. Dickey et al., *Russian Political Warfare: Origin, Evolution and Application*. Master's Thesis, Naval Postgraduate School, Monterey, CA, 2015.
95 Ibid.
96 O. Lutsevych, "Agents of the Russian worldproxy groups in the contested neighbourhood," *Chatham House – The Royal Institute of International Affairs Reports*, April 2016.
97 A. Applebaum, "Putin's grand strategy," *South Central Review* 35, no. 1 (Spring 2018).
98 A. Natsios, "Introduction: Putin's new Russia: Fragile state or revisionist power?" *South Central Review* 35, no. 1 (Spring 2018).
99 Applebaum, "Putin's grand strategy."
100 A. Polyakova, "Weapons of the weak: Russia and AI-driven asymmetric warfare," *Brookings Series: A Blueprint for the Future of AI: 2018-2019* (2018).
101 Q. Liu, "Important tasks to be fulfilled in the next five years," *China Daily Global Edition*, 6 March 2021; W. Callahan, "China 2035: From the China dream to the world dream," *Global Affairs* 2, no. 3 (2016).
102 U.S. Department of State, "The Chinese communist party: Threatening global peace and stability," (2021).
103 E. Crawford, "Made in China 2025: The industrial plan that China doesn't want anyone talking about," *PBS Online*, 7 May 2019.
104 Liu, "Important tasks."
105 Callahan, "China 2035."
106 Statistics Times, *World GDP Ranking 2021*.
107 Graphika, "Ants in a web: Deconstructing Guo Wengui's online whistleblower movement," (2021).
108 Defense Intelligence Agency (DIA), *China Military Power: Modernizing a Military Force to Fight and Win* (Washington, D.C., 2019).
109 J. Daniels, "Chinese theft of sensitive US military technology is still a 'huge problem,' says defense analyst," *CNBC*, 8 November 2017.
110 S. Roblin, "Why China's latest jets are surpassing Russia's top fighters," *Forbes*, 10 November 2020.
111 J. Huang, "Chinese drugmakers play catch up on mRNA vaccines amid pandemic," *S&P Global Market Intelligence*, 10 February 2021.
112 C. Bing and M. Taylor, "Exclusive: China-backed hackers 'targeted COVID-19 vaccine firm Moderna," *Reuters*, 30 July 2020.
113 D. Cave and J. Wallis, "Defending democracies from disinformation and cyber-enabled foreign interference in the COVID-19 era," *Observer Research Foundation Issue Briefs and Special Reports*, (12 April 2021).
114 K. Jamieson and D. Albarracin, "The relation between media consumption and misinformation at the outset of the SARS-CoV-2 pandemic in the US," *The Harvard Kennedy School (HKS) Misinformation Review* 1, no. 2 (2020).
115 P.I. Ryan, "Most Americans think they can spot fake news. They can't, study finds," *CNN Health*, 31 May 2021.
116 K. Gradon, "Crime in the time of the plague: Fake news pandemic and the challenges to law-enforcement and intelligence community," *Society Register* 4, no. 2 (2020).
117 I. Freckelton, "COVID-19: Fear, quackery, false representations and the law," *International Journal of Law and Psychiatry* 72 (2020).
118 S. Evanega et al., "Coronavirus misinformation: Quantifying sources and themes in the COVID-19 'infodemic'," *JMIR Preprints* 19, no. 10 (2020).
119 J. Brennen et al., *Types, Sources, and Claims of COVID-19 Misinformation* (Reuters Institute, 2020).

120 S. Bond, "Just 12 people are behind most vaccine hoaxes on social media, research shows," *Untangling Misinformation*, National Public Radio, 2021; Center for Countering Digital Hate, *Pandemic Profiteers: The Business of Anti-Vaxx* (2021).
121 J. Roozenbeek et al., "Susceptibility to misinformation about COVID-19 around the world," *Royal Society Open Science* 7 (2020).
122 T. Brewster, "Google promised not to use its AI in weapons, so why is it investing in startups straight out of 'star wars'?" *Forbes*, 22 December 2020.
123 T. Castelino, "Google renounces AI work on weapons," *Arms Control Today*, July 2018.
124 U.S. National Security Commission on Artificial Intelligence, *Final Report*.
125 The authors would like to acknowledge that information related to Artificial Intelligence was extracted from publicly available sources along with help and assistance from AI experts who wish to remain anonymous.

Bibliography

Abid, A., M. Farooqi, and J. Zou. "Persistent anti-Muslim bias in large language models." arXiv:2101.05783 [cs.CL] (2021).

Antonakaki, D., P. Fragopoulou, and S. Ioannidis. "A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks." *Expert Systems with Applications* 164 (2021): 114006.

Apetroe, A.C. "Hybrid warfare: From war during peace to neo-imperialist ambitions. The case of Russia." *On-line Journal Modelling the New Europe* 21, no. 1 (2016): 97–128.

Applebaum, A. "Putin's grand strategy." *South Central Review* 35, no. 1 (Spring 2018): 22–34. https://doi.org/10.1353/scr.2018.0001.

Aral, S., L. Muchnik, and A. Sundararajan. "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks." *Proceedings of the National Academy of Sciences* 106, no. 51 (2009): 21544–21549.

Artime, O., V. d'Andrea, R. Gallotti, P.L. Sacco, and M. De Domenico. "Effectiveness of dismantling strategies on moderated vs. unmoderated online social platforms." *Scientific Reports* 10, no. 1 (2020): 1–11.

Atkinson, C. "Hybrid warfare and societal resilience: Implications for democratic governance." *Information & Security* 39, no.1 (2018): 63–76. https://doi.org/10.11610/isij.3906.

Bail, C.A., B. Guay, E. Maloney, A. Combs, D. Sunshine Hillygus, F. Merhout, D. Freelon, and A. Volfovsky. "Assessing the Russian internet research agency's impact on the political attitudes and behaviors of American Twitter users in late 2017." *Proceedings of the National Academy of Sciences* 117, no. 1 (2020): 243–250.

Barabási, A.-L., and R. Albert. "Emergence of scaling in random networks." *Science* 286, no. 5439 (1999): 509–512.

Barton, R. "Tutorial: Simulation metamodeling." *2015 Winter Simulation Conference (WSC)* (2015): 1765–1779. https://doi.org/10.1109/WSC.2015.7408294.

Battiston, F., G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri. "Networks beyond pairwise interactions: structure and dynamics." *Physics Reports* 874 (2020): 1–92.

Bender, E.M., T. Gebru, A. McMillan-Major, and S. Schmitchell. "On the dangers of stochastic parrots: can language models be too big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623, New York, NY: Association for Computing Machinery, 2021.

Bener, A.B., B. Çağlayan, A.D. Henry, and P. Prałat. "Empirical models of social learning in a large evolving network." *PloS one* 11, no. 10 (2016): e0160307.

Biden, J., and M. Carpenter. "How to stand up to the kremlin: Defending democracy against its enemies." *Foreign Affairs* 87, no. 1 (January/February 2018). www.foreignaffairs.org/print/note/1121401.

Bing, C., and M. Taylor. "Exclusive: China-backed hackers 'targeted COVID-19 vaccine firm Moderna." *Reuters*, 30 July 2020. https://www.reuters.com/article/us-health-coronavirus-moderna-cyber-excl/exclusive-china-backed-hackers-targeted-covid-19-vaccine-firm-moderna-idUSKCN24V38M.

Boghardt, T. "Soviet bloc intelligence and its AIDS disinformation campaign." *Studies in Intelligence* 53, no. 4 (December 2009): 5–8.

Bolukbasi, T., K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." arXiv:1607.06520 [cs.CL] (2016).

Bonabeau, E. "Agent-based modelling: Methods and techniques for simulating human systems." *Proceedings of the National Academy of Sciences* 99, suppl. 3 (May 2002): 7280–7287. https://doi.org/10.1073/pnas.082080899.

Bond, S. "Just 12 People are behind most vaccine hoaxes on social media, research shows." *Untangling Misinformation* (2021). https://www.npr.org/2021/05/13/996570855/disinformation-dozen-test-facebooks-twitters-ability-to-curb-vaccine-hoaxes.

Braw, E. "Beware of bad samaritans." *Foreign Policy*, 30 March 2020. https://foreignpolicy.com/2020/03/30/russia-china-coronavirus-geopolitics/.

Brennen, J., F. Simon, P. Howard, and R. Nielsen. *Types, Sources, and Claims of COVID-19 Misinformation*. Reuters Institute, 2020. https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation.

Brewster, T. "Google promised not to use its AI in weapons, so why is it investing in startups straight out of 'star wars'?" *Forbes*, 22 December 2020. https://www.forbes.com/sites/thomasbrewster/2020/12/22/google-promised-not-to-use-its-ai-in-weapons-so-why-is-alphabet-investing-in-ai-satellite-startups-with-military-contracts/?sh=42794e787595.

Broniatowski, D., D. Kerchner, F. Farooq, X. Huang, A. Jamison, M. Dredze, and S. Crouse Quinn. "The covid-19 social media infodemic reflects uncertainty and state-sponsored propaganda." (2020). https://www.researchgate.net/profile/Xiaolei-Huang-3/publication/343096174_The_COVID-19_Social_Media_Infodemic_Reflects_Uncertainty_and_State-Sponsored_Propaganda/links/5f4a8beca6fdcc14c5e25393/The-COVID-19-Social-Media-Infodemic-Reflects-Uncertainty-and-State-Sponsored-Propaganda.pdf.

Brown, T.B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al. "Language models are few-shot learners." arXiv:2005.14165 [cs.CL] (2020).

Byrne, D. "An overview (and underview) of research and theory within the attraction paradigm." *Journal of Social and Personal Relationships* 14, no. 3 (1997): 417–431.

Callahan, W. "China 2035: From the China dream to the world dream." *Global Affairs* 2, no. 3 (2016): 247–258.

Carlini, N., F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts et al. "Extracting training data from large language models." arXiv:2012.07805 [cs.CR] (2020).

Carmi, E., S.J. Yates, E. Lockley, and A. Pawluczuk. "Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation." *Internet Policy Review* 9, no. 2 (2020). https://doi.org/10.14763/2020.2.1481.

Carter, D. "Hustle and brand: The sociotechnical shaping of influence." *Social Media and Society* 2, no. 3 (2016).

Castelino, T. "Google renounces AI work on weapons." *Arms Control Today*, July 2018. https://www.armscontrol.org/act/2018-07/news/google-renounces-ai-work-weapons.

Cave, D., and J. Wallis. "Defending democracies from disinformation and cyber-enabled foreign interference in the COVID-19 era." *Observer Research Foundation Issue Briefs and Special Reports*, 12 April 2021. https://www.orfonline.org/research/defending-democracies-from-disinformation-and-cyber-enabled-foreign-interference-in-the-covid-19-era/.

Center for Countering Digital Hate. *Pandemic Profiteers: The Business of Anti-Vaxx*, 2021. www.counterhate.org.

Cha, M., H. Haddadi, F. Benevenuto, and K. Gummadi. "Measuring user influence in Twitter: The million follower fallacy." In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, no. 1, 2010.

Chen, T., X. Li, H. Yin, and J. Zhang. "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection." In *Pacific-Asia conference on knowledge-discovery and data mining*, edited by M. Ganji, L. Rashidi, B. Fung and C. Wang, 40–52, Springer, 2018.

Crawford, E. "Made in China 2025: The industrial plan that china doesn't want anyone talking about." *PBS Online*, 7 May 2019. https://www.pbs.org/wgbh/frontline/article/made-in-china-2025-the-industrial-plan-that-china-doesnt-want-anyone-talking-about/.

Cunningham, R. "Introducing power apps ideas: AI-powered assistance now helps anyone create apps using natural language." *Microsoft Power Apps* blog, 25 May 2021. https://powerapps.microsoft.com/nl-nl/blog/introducing-power-apps-ideas-ai-powered-assistance-now-helps-anyone-create-apps-using-natural-language/.

Daniels, J. "Chinese theft of sensitive US military technology is still a 'huge problem,' says defense analyst." *CNBC*, 8 November 2017. https://www.cnbc.com/2017/11/08/chinese-theft-of-sensitive-us-military-technology-still-huge-problem.html.

De Domenico, M., and E.G. Altmann. "Unraveling the origin of social bursts in collective attention." *Scientific Reports* 10, no. 1 (2020): 1–9.

Defense Intelligence Agency (DIA). *China Military Power: Modernizing a Military Force to Fight and Win*. Washington, DC, 2019. https://www.dia.mil/Portals/27/Documents/News/Military%20Power%20Publications/China_Military_Power_FINAL_5MB_20190103.pdf.

Dickey, J.V., T.B. Everett, Z.M. Galvach, M.J. Mesko, and A.V. Soltis. *Russian Political Warfare: Origin, Evolution and Application*. Master's Thesis, Naval Postgraduate School, Monterey, CA, 2015. https://www.hsdl.org/?view&did=811550.

Emmott, R. "Russia, China sow disinformation to undermine trust in Western vaccines: EU." *Reuters*, 28 April 2021. https://www.reuters.com/world/china/russia-china-sow-disinformation-undermine-trust-western-vaccines-eu-report-says-2021-04-28/.

Evanega, S., M. Lynas, J. Adams, K. Smolenyak, and C.G. Insights. "Coronavirus misinformation: Quantifying sources and themes in the COVID-19 'infodemic'." *JMIR Preprints* 19, no. 10 (2020).

Faloutsos, M., P. Faloutsos, and C. Faloutsos. "On power-law relationships of the internet topology." In *The Structure and Dynamics of Networks*, Princeton, NJ: Princeton University Press, 2011.

Feng, Y., H. You, Z. Zhang, R. Ji, and Y. Gao. "Hypergraph neural networks." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, no. 1 (2019): 3358–3565.

Freckelton, I. "COVID-19: Fear, quackery, false representations and the law." *International Journal of Law and Psychiatry* 72 (2020): 101611.

Freelon, D., M. Bossetta, C. Wells, J. Lukito, Y. Xia, and K. Adams. "Black trolls matter: Racial and ideological asymmetries in social media disinformation." *Social Science Computer Review* 89 (2020): 148–153.

Frye, T. "Russia's weak strongman: The perilous bargains that keep Putin in power." *Foreign Affairs* 100, no. 3 (2021).

Gallegati, M., and M.G. Richiardi. "Agent based models in economics and complexity." In *Encyclopedia of Complexity and Systems Science*, edited by R. Meyers. New York: Springer, 2009. https://doi.org/10.1007/978-0-387-30440-3.

Gallotti, R., F. Valle, N. Castaldo, P. Sacco, and M. De Domenico. "Assessing the risks of 'infodemics' in response to COVID-19 epidemics." *Nature Human Behaviour* 4, no. 12 (2020): 1285–1293.

Gilbert, N. *Agent-Based Models*. SAGE, 2008.

Gleeson, J.P., K.P. O'Sullivan, R.A. Baños, and Y. Moreno. "Effects of network structure, competition, and memory time on social spreading phenomena." *Physical Review* X6, no. 2 (2016): 021019.

González-Bailón, S., and M. De Domenico. "Bots are less central than verified accounts during contentious political events." *Proceedings of the National Academy of Sciences* 118, no.11 (2021).

Gradon, K. "Crime in the time of the plague: Fake news pandemic and the challenges to law-enforcement and intelligence community." *Society Register* 4, no. 2 (2020). https://pressto.amu.edu.pl/index.php/sr/article/view/22513.

Gradon, K., J.A. Hołyst, W.R. Moy, J. Sienkiewicz, and K. Suchecki. "Countering misinformation: A multidisciplinary approach." *Big Data & Society Special Issue on Studying Infodemic at Scale* 89, no. 1 (May 2021). https://doi.org/10.1177/20539517211013848.

Graphika. "Ants in a web: Deconstructing Guo Wengui's online whistleblower movement." (17 May 2021). https://www.graphika.com/reports/ants-in-a-web.

Guo, B., Y. Ding, L. Yao, Y. Liang, and Z. Yu. "The future of false information detection on social media: new perspectives and trends." *ACM Computing Surveys (CSUR)* 53, no. 4 (2020): 1–36.

Guo, W., and A. Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." arXiv:2006.03955 [cs.CY] (2020).

Hamilton, W.L., R. Ying, and J. Leskovec. "Representation learning on graphs: Methods and applications." arXiv:1709.05584 [cs.SI] (2017).

Hébert-Dufresne, L., S.V. Scarpino, and J.-G. Young. "Macroscopic patterns of interacting contagions are indistinguishable from social reinforcement." *Nature Physics* 16, no. 4 (2020): 426–431.

Henry, A.D., P. Prałat, and C.-Q. Zhang. "Emergence of segregation in evolving social networks." *Proceedings of the National Academy of Sciences* 108, no. 21 (2011): 8605–8610.

Himelboim, I., S. McCreery, and M. Smith. "Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter." *Journal of Computer-Mediated Communication* 18, no. 2 (2013): 154–174.

Huang, J. "Chinese drugmakers play catch up on mRNA vaccines amid pandemic." *S&P Global Market Intelligence*, 10 February 2021. https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/chinese-drugmakers-play-catch-up-on-mrna-vaccines-amid-pandemic-62527322.

Jamieson, K., and D. Albarracin. "The relation between media consumption and misinformation at the outset of the SARS-CoV-2 pandemic in the US." *The Harvard Kennedy School (HKS) Misinformation Review* 1, no. 2 (2020). https://doi.org/10.37016/mr-2020-012.

Jawahar, G., M. Abdul-Mageed, and L.V.S. Lakshmanan. "Automatic detection of machine generated text: A critical survey." arXiv:2011.01314 [cs.CL] (2020).

Jin, Z., J. Cao, Y. Zhang, J. Zhou, and Q. Tian. "Novel visual and statistical image features for microblogs news verification." *IEEE Transactions on Multimedia* 19, no. 3 (2016): 598–608.

Kamínski, B. "A method for the updating of stochastic kriging metamodels." *European Journal of Operational Research* 247, no. 3 (2015): 859–866.

Kamínski, B. "Interval metamodels for the analysis of simulation Input-Output relations." *Simulation Modelling Practice and Theory* 54 (2015): 86–100.

Kamínski, B., V. Poulin, P. Prałat, P. Szufel, and F. Théberge. "Clustering via hypergraph modularity." *PloS one* 14, no. 11 (2019) e0224307.

Kamínski, B., P. Prałat, and F. Théberge. "An unsupervised framework for comparing graph embeddings." *Journal of Complex Networks* 8, no. 5 (2020): cnz043.

Kamínski, B., P. Prałat, and F. Théberge. *Mining Complex Networks*. New York: CRC Press, 2021.

Kaneko, M., and D. Bollegala. "Gender-preserving debiasing for pre-trained word embeddings." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1641–1650, Florence: Association for Computational Linguistics, 2019.

Kasapoglu, C., and B. Kirdemir. "Artificial intelligence and the future of conflict." In *New Perspectives on Shared Security: NATO's Next 70 Years*, edited by T. Valasek, Brussels: Carnegie Europe, 2019. https://carnegieendowment.org/files/NATO_int_final1.pdf.

Kinetz, E. "COVID conspiracy shows vast reach of Chinese disinformation." *The Associated Press*, 15 February 2021. https://apnews.com/article/beijing-media-coronavirus-pandemic-conspiracy-only-on-ap-e696b32d4c3e9962ac0bdbdae2991466.

Kirk, H., Y. Jun, H. Iqbal, E. Benussi, F. Volpin, F.A. Dreyer, A. Shtedritski, and Y.M. Asano. "How true is GPT-2? An empirical analysis of intersectional occupational biases." (2021). https://arxiv.org/pdf/2102.04130v1.pdf.

Kucharski, A. *The Rules of Contagion: Why Things Spread – and Why They Stop*. New York: New York Basic Books, 2020.

Kumar, A., A. Biswas, and S. Sanyal. "eCommerceGAN: A generative adversarial network for e-commerce." arXiv:1801.03244 [cs.LG] (2018).

Kusner, M.J., and J.M. Hernández-Lobato. "Gans for sequences of discrete elements with the gumbel-softmax distribution." arXiv:1611.04051 [stat.ML] (2016).

Kwon, S., M. Cha, and K. Jung. "Rumor detection over varying time windows." *PloS One* 12, no. 1 (2017): e0168344.

Lehmann, J., B. Gonçalves, J.J. Ramasco, and C. Cattuto. "Dynamical classes of collective attention in Twitter." In *Proceedings of the 21st International Conference on World Wide Web*, New York: Association for Computing Machinery, 2012.

Leskovec, J., J., Kleinberg, and C. Faloutsos. "Graph evolution: Densification and shrinking diameters." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, no. 1 (2007): 2-es.

Liu, Q. "Important tasks to be fulfilled in the next five years." *China Daily Global Edition*, 6 March 2021. https://global.chinadaily.com.cn/a/202103/06/WS6042d816a31024ad0baad375.html.

Liu, Y., and Y.-F.B. Wu. "Fned: A deep network for fake news early detection on social media." *ACM Transactions on Information Systems (TOIS)* 38, no. 3 (2020): 1–33.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy et al. "RoBERTa: A robustly optimized BERT pretraining approach." arXiv:1907.11692 [cs.CL] (2019).

Lutsevych, O. "Agents of the Russian worldproxy groups in the contested neighbourhood." *Chatham House – The Royal Institute of International Affairs Reports*, April 2016. https://www.chathamhouse.org/sites/default/files/publications/research/2016-04-14-agents-russian-world-lutsevych.pdf.

Ma, J., W. Gao, P. Mitra, S. Kwon, B.J. Jansen, K.-F. Wong, and M. Cha. "Detecting rumors from microblogs with recurrent neural networks." *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, edited by Subbarao Kambhampati, AAAI Press / International Joint Conferences on Artificial Intelligence, 2016.

Manzini, T., Y.C. Lim, Y. Tsvetkov, and A.W. Black. "Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings." arXiv:1904.04047 [cs.CL] (2019).

McGuffie, K., and A. Newhouse. "The radicalization risks of GPT-3 and advanced neural language models." arXiv:2009.06807 [cs.CY] (2020).

Moy, W.R., and K. Gradon. "COVID-19 effects and Russian disinformation." *Homeland Security Affairs* 16 (2020): 8.

Munteanu, R. "Hybrid Warfare: The new form of conflict at the beginning of the century." *Strategic Impact* 56, no. 3 (2015): 19.

Natsios, A. "Introduction: Putin's New Russia: Fragile state or revisionist power?" *South Central Review* 35, no. 1 (Spring 2018): 1–21. https://doi.org/10.1353/scr.2018.0000.

Nguyen, V.-H., K. Sugiyama, P. Nakov, and M.-Y. Kan. "FANG: Leveraging social context for fake news detection using graph representation." *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020. https://arxiv.org/pdf/2008.07939.pdf.

Polyakova, A. "Weapons of the weak: Russia and AI-driven asymmetric warfare." *Brookings Series: A Blueprint for the Future of AI: 2018-2019* (2018) https://www.brookings.edu/research/weapons-of-the-weak-russia-and-ai-driven-asymmetric-warfare/.

Raj, P., and E. Preetha. *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases*, Elsevier, 2020.

Rajeswar, S., S. Subramanian, F. Dutil, C. Pal, and A. Courville. "Adversarial generation of natural language." arXiv:1705.10929 [cs.CL] (2017).

Roblin, S. "Why China's latest jets are surpassing Russia's top fighters." *Forbes*, 10 November 2020. https://www.forbes.com/sites/sebastienroblin/2020/11/10/why-chinas-latest-jets-are-surpassing-russias-top-fighters/?sh=7b50ef382e26.

Roozenbeek, J., C.R. Schneider, S. Dryhurst, J. Kerr, A.L.J. Freeman, G. Recchia, A.M. van der Bles, and S. van der Linden. "Susceptibility to misinformation about COVID-19 around the world." *Royal Society Open Science* 7 (2020): 2011199. http://dx.doi.org/10.1098/rsos.201199.

Rothkopf, D.J. "When the buzz bites back." *The Washington Post*, 11 May 2003. https://www.washingtonpost.com/archive/opinions/2003/05/11/when-the-buzz-bites-back/bc8cd84f-cab6-4648-bf58-0277261af6cd/.

Rothman, D. *Hands-On Explainable AI (XAI) with Python*. O'Reilly, 2020.

Ruchansky, N., S. Seo, and Y. Liu. "Csi: A hybrid deep model for fake news detection." *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, 2017. https://doi.org/10.48550/arXiv.1703.06959.

Ryan, P.I. "Most Americans think they can spot fake news. They can't, study finds." *CNN Health*, 31 May 2021. https://www.cnn.com/2021/05/31/health/fake-news-study/index.html.

Schelling. T. "Models of segregation." *The American Economic Review* 59, no. 2 (1969): 488–493.

Schnaufer, T.A. "Redefining hybrid warfare: Russia's nonlinear war against the west." *Journal of Strategic Security* 10, no. 1 (2017): 17–31. https://doi.org/10.5038/1944-0472.10.1.1538.

Schramowski, P., C. Turan, N. Andersen, C. Rothkopf, and K. Kersting. "Language models have a moral dimension." Preprint. (2021). https://www.researchgate.net/publication/350310986_Language_Models_have_a_Moral_Dimension.

Schuster, T., R. Schuster, D.J. Shah, and R. Barzilay. "The limitations of stylometry for detecting machine-generated fake news." *Computational Linguistics* 46, no. 2 (2020): 499–510.

Shu, K., A. Sliva, S. Wang, J. Tang, and H. Liu. "Fake news detection on social media: A data mining perspective." *ACM SIGKDD Explorations Newsletter* 19, no. 1 (2017): 22–36.

Solaiman, I., M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford. "Release strategies and the social impacts of language models." arXiv:1908.09203 [cs.CL] (2019).

Statistics Times. *World GDP Ranking 2021*, 2021. https://statisticstimes.com/economy/projected-world-gdp-ranking.php.

Stella, M., M. Cristoforetti, and M. De Domenico. "Influence of augmented humans in online interactions during voting events." *PloS One* 14, no. 5 (2019): e0214210.

Tamkin, A., M. Brundage, J. Clark, and D. Ganguli. "Understanding the capabilities, limitations, and societal impact of large language models." arXiv:2102.02503 [cs.CL] (2021).

Turcilo, L., and M. Obrenovic. "Misinformation, disinformation, malinformation: Causes, trends, and their influence on democracy." *Heinrich Böll Foundation Companion to Democracy,* 3 August 2020. https://hk.boell.org/sites/default/files/importedFiles/2020/11/04/200825_E-Paper3_ENG.pdf.

U.S. Department of State. "The Chinese communist party: Threatening global peace and stability." (2021) https://www.state.gov/wp-content/uploads/2020/10/FINAL20one-pager20Threatening20Global20Peace20Security-1.pdf.

U.S. National Security Commission on Artificial Intelligence. *Final Report*. 2021. https://www.nscai.gov/.

Wardle, C., and H. Derakhshan. "Information disorder: Towards an interdisciplinary framework for research and policy-making." *Council of Europe Report*, 27 September 2017. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

Watts, D.J., and S.H. Strogatz. "Collective dynamics of 'small-world' networks." *Nature* 393, no. 6684 (1998): 440–442.

Will, M., J. Groeneveld, K. Frank, and B. Müller. "Combining social network analysis and agent-based modelling to explore dynamics of human interaction: A review." *Socio-Environmental Systems Modelling* 2 (2020). https://doi.org/10.18174/sesmo.2020a16325.

Williams, R.J. "Simple statistical gradient-following algorithms for connectionist reinforcement learning." *Machine Learning* 8, no. 3–4 (1992): 229–256.

Woolley, S.C., and P.N. Howard. *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. New York: Oxford University Press, 2018.

World Health Organization (WHO). "Coronavirus disease 2019 (COVID-19)." *Situation Report* 45, 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200305-sitrep-45-covid-19.pdf.

World Health Organization (WHO). *1st WHO Infodemiology Conference: How Infodemics Affect the World & How They can be Managed*. Conference Booklet, 2020. https://www.who.int/docs/default-source/epi-win/infodemic-management/infodemiology-scientific-conference-booklet.pdf?sfvrsn=179de76a.

Wu, L., and H. Liu. "Tracing fake-news footprints: Characterizing social media messages by how they propagate." *Proceedings of the eleventh ACM International Conference on Web Search and Data Mining*, 637–645, New York: Association for Computing Machinery, 2018.

Zellers, R., A. Holtzman, H. Rashkin. Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. "Defending against neural fake news." arXiv:1905.12616 [cs.CL] (2019).

Zhang, D., J. Yin, X. Zhu, and C. Zhang. "Network representation learning: A survey." *IEEE Transactions on Big Data* 6, no. 1 (2018): 3–28.

Zhou, X., A. Jain, V.V. Phoha, and R. Zafarani. "Fake news early detection: A theory-driven model." *Digital Threats: Research and Practice* 1, no. 2 (2020): 1–25.

Part II

Strategic and geopolitical challenges

4 Algorithmic power?

The role of artificial intelligence in European strategic autonomy

Simona R. Soare

Introduction

The European Union and its member states are investing in a wide range of emerging and disruptive technologies[1] (EDTs) to be the backbone of European strategic autonomy. The Sino-American strategic and technological competition is pushing Europe into a geopolitically tight space. Brussels fears the geopolitical and geoeconomic consequences of lagging behind the US and China, France fears "irreversible dependencies"[2] and Germany fears the loss of European geoeconomic competitiveness[3] if Europeans do not strengthen their technological sovereignty and strategic autonomy.

The US wants Europe to invest more in defence and safeguard against malign Chinese investments in key technologies and critical infrastructure. However, Washington remains sceptical of European strategic autonomy for geopolitical reasons, particularly if it leads to a European "third way" or European "equidistance" between the US and China.[4] Beijing is encouraging European strategic autonomy notably from the US, and Russia is sceptically weighing the alternative futures of the European project.[5] In short, international actors do not see sufficient strategic intent and technological and hard power behind the EU's narrative on strategic autonomy. As demonstrated by the 5G debate, the US and China perceive Europe as the grounds for geopolitical confrontation – an outcome Brussels and major European capitals badly want to avoid.

Investing in European-grown and controlled EDTs (including supporting data and digital infrastructure) and defence capabilities has become a European priority linked to strategic autonomy. And the COVID-19 pandemic has accelerated European strategic planning in this regard. Technology is key to both self-perceptions and international perceptions of the EU status as one of the leading geopolitical players.[6] Strategic autonomy requires a degree of differentiation from and sustainable competitiveness with other great powers, notably the US and China.[7] Arguably, this is the impetus behind what European Commission president von der Leyen calls "a digital transition which is *European by design and nature*."

What role do AI technologies play in European strategic autonomy in security and defence? At present, the EU is ill-equipped to leverage AI

DOI: 10.4324/9781003284093-6

technologies in defence and transfer them to the battlefield in ways that lead to greater strategic autonomy in defence. The paper argues insofar as European strategic autonomy is concerned, the adoption of AI technologies in defence is a distraction because it introduces new layers of complexity to European defence without significantly contributing to greater European strategic autonomy.

Four interrelated reasons explain this. First, Europeans lack *a common AI integration strategy in defence* which links technological power to strategic autonomy in terms of operational advantage against and competitiveness with other rival great powers. The EU does not have, nor does it currently plan to develop a *common European military strategy to integrate AI in cyber and cross-domain military operations* for operational and strategic advantage and it does not possess *a common, regular threat and opportunity assessment based on European intelligence* about its rivals' AI military innovation efforts and other international actors' geopolitical needs. The EU is not politically ready – or interested – to develop the kind of military capabilities, enablers, and legal powers to conduct algorithmic warfare.

Second, *Europeans are still struggling to close important technology and capability gaps*, including in the fields of cyber and digital emerging technologies. However, as cyber, and digital technologies become the backbone of cutting-edge military capabilities and shift the global balance of power, greater strategic autonomy depends on European mastery of these new forms of power as effective tools of military statecraft. So far, the EU's approach to strategic autonomy and technology lacks a clear geopolitical and proactive focus, and resides firmly in a normative, regulatory, and reactive dimension that is devoid of a strong technological and hard power foundation.

Third, *EU decision-making in foreign, security and defence policy is ill-equipped to accommodate the use of AI technologies*. There is currently no clear understanding of what role AI would play in supporting EU decision-making and of the differentiated information needs from AI systems at different levels of decision-making, across the civil-military spectrum. Ongoing debates about Qualified Majority Voting (QMV) and article 44 TEU do not offer an adequate solution because they do not reflect deeper adaptations needed to accommodate and integrate the use of AI. Insofar as AI can deliver information and not political consensus, it is arguable whether AI technologies are an adequate cure for Europe's 'bias for inaction' which stems from a fundamental lack of political will to act, common threat assessment and agreement on how to tackle these threats, including through the use of force.

Fourth, *the role of AI technologies in enabling greater European strategic autonomy is secondary to geopolitical and exogenous factors*. Geopolitical and structural factors such as great power competition and threat perception, the structure and functioning of the global internet, global AI standards, strategic supply chains, European relations to private industry, interdependence and global geopolitical alignment patterns are powerful constraints that affect European strategic autonomy regardless of the availability of European technologies.

This paper seeks to break new ground by providing a conceptual and empirical analysis of the relationship between technology and political autonomy in international organisations, alliances, and multilateral fora. It contributes to security and technology literature by conceptualising the relationship between AI technologies, the shift in the distribution of global power and political constructs of power and autonomy. It does so by framing European strategic autonomy as a strategy for great power competition at four levels: decision-making, technological, operational autonomy, and partnerships. The paper also highlights the pitfalls of techno-centric solutions and debunks frequent assumptions about the equivalency between mastering new technologies and employing them as effective tools of statecraft in strategic competitions.

The paper proceeds in three sections. Section one explores the link between European strategic autonomy, AI, and cyber power. Section two offers a radiography of the EU's efforts and strategy development towards AI adoption and assesses their early results. The third section analyses European efforts to adopt AI in defence in relation to European decision-making autonomy, technological autonomy, operational autonomy, and defence partnerships, notably EU-NATO cooperation.

Algorithmic power and strategic autonomy

A strategy for international competition?

Since 2016, the key EU objective has been to achieve a generically defined "appropriate level of ambition and strategic autonomy"[8] "while preserving an open economy."[9] The European Council defined strategic autonomy as the "capacity to act autonomously when and where necessary and with partners wherever possible."[10] The concept encompasses four main dimensions: decision-making autonomy, technological autonomy, operational autonomy,[11] and partnerships.

European strategic autonomy has developed and has been used in a European defence and defence industrial context,[12] in relation to the European development, possession, projection, and employment of hard power. The COVID-19 pandemic and the rapid changes in the international system[13] contributed to the broadening and refinement of the meaning of strategic autonomy as well as to its prioritisation.[14] President of the Council, Charles Michel reflected this trend by defining strategic autonomy as "more resilience, more influence. And less dependence" on other international actors. Strategic autonomy is about "being able to make choices" by "reducing our dependencies, to better defend our interests and our values."[15]

According to this broader understanding, strategic autonomy is the EU's long-term strategy[16] for international great power competition. It is driven by the recognition of a shifting global balance of power, the development of autonomous decision-making capacities and capabilities (political,

institutional, and material; soft and hard power), the reduction of critical external dependencies, and the emulation[17] of other great powers' geopolitical behaviour. This is further confirmed by the apparent conceptual overlap with another concept the EU and European states use in parallel, namely "strategic sovereignty."[18]

Both strategic autonomy and strategic sovereignty encompass internal and external dimensions. They concern: deepening integration (possibly towards an implied federalist project denoted by the use of the concept "sovereignty") and the full use of EU treaty provisions; domestic resilience and the reduction of strategic dependencies on external actors, securing critical supply chains and controlling critical infrastructure; the EU's internal capacity to command and control critical technologies which are key to its economic interests and security; and the EU's capabilities and political will to act externally to defend its strategic interests.[19] Both concepts seek to empower the EU *and* the member states acting collectively,[20] as exemplified by the expressed desire for more flexibility between CSDP and *ad hoc* missions of the member states in the Strategic Compass negotiations.[21] They also signal an expectation of *international recognition* by other great powers of the EU and the Europeans' peer status as credible security providers.

Weak links

Nevertheless, European military strategic autonomy lacks credibility.[22] Europeans lack credible military capabilities and strategic enablers[23] in sufficient numbers and levels of readiness, interoperability, and digitalisation for modern warfare.[24] CSDP lacks a credible military level of ambition[25] as do different European minilateral cooperation formats. The Helsinki Headline Goals for a corps-level EU Rapid Reaction Force, deployable in 60 days and fully sustainable in the battlefield for up to a year, were never met and, in practice the level of ambition has been much lower.[26] The EU Battlegroups were created, but they were never operationally deployed.

Furthermore, digital technologies in and beyond the military domain represent an EU strategic vulnerability. While Europeans have been making progress in AI, this progress varies widely across AI technology stacks. Europe's "share of global patent applications in Big Data is the smallest among all Advanced Technologies," no European chip manufacturer produces competitive, cutting-edge semiconductors[27] for advanced AI and QC (Quantum Computing), and "the largest European cloud service provider accounts for less than 1% of total revenues generated in the European market."[28] Between 2009 and 2019, leading European states such as France and Germany relied 73% and 64% respectively on national investors for the funding of tech start-ups. The rest of the funding came mainly from the US and the UK.[29]

Since the late 2000s, European strategic dependency on rare earths and raw materials necessary for digital technologies and defence capabilities has increased. For many rare earths elements such as antinomy, which are used in

electronic-optical systems and semiconductors, Europe relies on Chinese suppliers and faces critical supply chain choke points.[30] Europe equally depends on limited foreign suppliers for digital components and semiconductors.[31]

Both security and defence, and digital technologies remain weak links in the EU's quest for strategic autonomy and status recognition as a geopolitical actor.

Creating European AI leadership

Progressive refinement of AI goals

Building on the priorities of the EU Strategy on Artificial Intelligence (2018),[32] the 2021 review of the *Coordinated Plan on Artificial Intelligence* marks a higher EU AI level of ambition by asserting the EU's qualitative AI leadership.[33] Furthermore, the 2020 *White Paper on Artificial intelligence* focuses on enhancing European AI *cooperation* through cross-border networks of excellence and pipelines of innovation as a means of reducing the significant AI policy, technological capacity, and output fragmentation between European states. The document seeks to achieve this by providing options for regulatory frameworks to support AI adoption and/or mitigate against legal and ethical risks from the use of these technologies.

Finally, the 2021 *Artificial Intelligence Act* is the EU's effort to shape the (global) standards of AI and assert its AI qualitative leadership. Following in the footsteps of the Global Data Protection Regulation (GDPR), this is perceived by the Commission as the materialisation of the EU's first mover advantage to establish a balanced approach between sufficient regulation and facilitating innovation, between security and civil liberties.

Cross-policy synergies

The EU seeks to create synergies and correlations between different industrial sectors, including incentivise spin-off and spin-in[34] effects by establishing cooperative frameworks (e.g., defence innovation networks, and innovation accelerators and incubators), promoting hybrid civil-defence technology standards, reducing duplication of effort and funding, and better coordinating European-based R&D&I and procurement, through the *Digital Compass* (2021)[35] and the *Action Plan on Synergies between the civil, defence and space industries* (2021). The Action Plan proposes three steps: first, to identify critical technologies and create an EU Observatory of Critical Technologies to "provide regular monitoring and analysis of critical technologies, their potential applications, value chains, needed research and testing infrastructure, desired level of EU control over them, and existing gaps and dependencies." Second, based on these classified reports, to develop and implement technology roadmaps that serve as the basis for concrete EU actions, capabilities, and decision-making processes that enhance its overall technological sovereignty.

And third, to pursue flagship projects that maximise European and cross-sector technological cooperation.[36]

There are indications that EU digital policies have matured over the past five years. The Union has established and updated policies in all critical sectors for AI adoption, including semiconductors, cyber space, data and digital networks and infrastructure, including 5G and cloud. The 2020 *EU Cybersecurity Policy* acknowledges the role of AI technologies in cyber defence and deterrence across the civilian and defence domains as well as the role cybersecurity plays in securing AI, the Internet of Things, and other emerging technologies.[37] In 2021 the European Union Agency for Cybersecurity (ENISA) published an AI threat landscape study outlining the main threat vectors and threat actors and proposed the establishment of "an AI toolbox (…) with concrete mitigation measures" for the threats to European AI systems.[38] However, the EU is hard pressed to find members that would contribute offensive cyber capabilities to its CSDP missions or to defend EU AI models.

A broad public mandate

The focus and acceleration of EU AI efforts are much needed in view of European strategic vulnerability and high public "digital distrust" and techno-nationalist tendencies in leading European countries. European publics and expert communities reflect these techno-nationalist tendencies: "trust in digital technology does not extend far beyond national borders: Europeans are sceptical of governments and companies from other European countries. This "digital distrust" is only exacerbated when looking across the Atlantic" and vis-à-vis China.[39] Recent surveys suggest a growing European public concern with the societal and security impact of AI, automation, and other modern weapon systems[40] and national dependencies on Chinese (54%), American (50%), and other EU (42%) digital technologies.

Early results of European AI efforts

As a result of Brussels' market structuring efforts, by July 2021, 20 EU member states and all four EFTA associated countries had adopted AI national strategies or plans.[41] While the EU and its member states continue to lag behind the US and China in AI, some reports suggest "the EU is catching up fast" and "AI is the most dynamically developing technology in terms of patent filing and start-up activity in Europe (outpacing the US)."[42] The focus on AI strategy development marked an increase in pledged national and EU-wide spending on digital and AI technologies. The EU committed to invest €20bn annually in AI for Europe's Digital Decade (2020–2030), of which €1bn annually are allocated through the Digital Europe and Horizon Europe programmes. In addition, €134bn are allocated to digital transformation under the *Recovery and Resilience Facility*. This European increase in AI funding is driven by domestic need and reduction of strategic dependencies as

well as by the perception of a trend among leading great powers to subsidise technological (and particularly AI) adoption across their societies. The Commission and European states are also slowly beginning to create and diversify innovative funding alternatives to accelerate AI adoption, including venture capital funding.[43]

Since 2017–2018, there is visible EU progress in the field of AI strategy development and funding. To the degree a European approach to AI is emerging, it is primarily normative and regulatory, bordering on variable degrees of techno-nationalism, Eurocentrism, and European exceptionalism.[44] It is driven by the ambition of asserting a qualitative European leadership in technology matters at the global level by exploiting first mover market integration, regulatory and hybrid standardisation effects. And it is marked by uneven progress between the civilian and the security and defence fields, closely reflecting the status of the broader EU integration process.

AI and European defence

There are great differences between European states with regard to AI adoption in defence. As recently emphasised by EU officials, the status quo "is not sufficient if only one or two Member States have [military AI applications]. They must be available across all EU Member States."[45] EU adoption of AI in defence has so far concentrated on three main lines of effort which are not always well coordinated.

Political efforts

The first line of effort is *political*. It resides within the Council, the European Parliament and, to a lesser extent, the External Action Service.[46] Rather than facilitating a strategic common EU policy approach to AI in defence, different political processes remain siloed and fragmented. The first European high-level exploratory consultation of digitalisation and AI in defence took place in August 2019, during the Finnish presidency of the Council and incentivised two parallel processes: an EDA process to raise awareness on AI in defence and an EU Military Staff process to accelerate defence modernisation through the digitalisation of defence. In 2020, the German presidency proposed an EU process on the 'Responsible military use of artificial intelligence' to agree a set of common norms and standards for the operational use of AI and the export of AI-enabled military systems. However, this process is not synchronised with any other ongoing AI work in the EU and has not yielded any results given the opposition of other member states.[47] The European Parliament has explored the manifold role of AI in defence[48] and external action.[49] However, this has done little to change its political position driven by a normative logic of AI adoption in defence rather than one focused on shaping AI adoption towards achieving greater military advantage.

The Parliament continues to insist on strict regulation and even ban of AI-enabled autonomy in defence, including in cyber defence.[50]

Technical efforts

The second line of EU effort is *technical*. It is concentrated within the European Commission, the EU Military Staff and the European Defence Agency and fostered within the EU defence initiatives,[51] though there is no coherent strategy linking the capabilities developed in Permanent Structured Cooperation (PESCO) and European Defence Fund (EDF) projects to specific EU operational doctrine and clear metrics of greater operational capacity and interoperability necessary for greater strategic autonomy.[52] The publication of the 2021 EDF work program suggests greater European emphasis on AI and other EDTs,[53] which makes a common European approach an urgent necessity.

The narrative around several loosely connected EU work strands promises to deliver such an approach. A 2021 EEAS Memo on the status of the *Strategic Compass* negotiations lists "strengthen[ing] the European Technological and Industrial Base including in particular an enhanced common EU approach to emerging and disruptive technologies in the security and defence domain" as one of three priorities in the 'Capabilities basket.'[54] However, the Strategic Compass document generically mentions emerging technologies in the context of long-term, next-generation capabilities investment.

In 2021, the European Commission worked on a *Roadmap on Critical Technologies for Security and Defence* which operationalised the *Action Plan on Synergies between Civil, Defence and Space Industries*, notably by identifying means to incentivise and support civilian innovation with potential defence applications. The *Roadmap*, published in 2022, guides further Commission action on technological roadmaps, in conjunction with the Commission's EDTs Observatory. However, even the *Roadmap*'s proponents admittedly do not have a long-term perspective on the utility of this tool.[55]

Since 2019, the EDA has developed an *AI in Defence Definition, Taxonomy and Glossary*, an *AI in Defence Narrative*, and *AI in Defence Action Plan* which could be linked to capability development and collaborative European R&T projects. Based on these, the EDA has developed an AI Strategic Research Agenda and identified over 50 technology building blocks relevant to AI. And in 2021, the EDA started preparing an *EDTs Action Plan* which will support EU states in "monitoring the EDT landscape in and outside of the EU" through horizon scanning and foresight and identifying and pursuing collaborative R&T projects "to avoid fragmentation and duplication."[56]

The EDA's 2020 Coordinated Annual Review on Defence (CARD) report indicates EU defence initiatives are too recent to effectively steer national defence planning and it is worth asking whether empirical evidence supports the logic behind capability development in the EU, from the CDP

and OSRA to CARD, EDF and PESCO, in the context of military AI. However, the lack of convergence between different EU innovation initiatives is not just a symptom of intensifying inter-agency power struggles to expand competencies and political clout. They are also evidence that a common AI approach in defence is yet again failing to emerge among key EU institutions and member states.

Operational efforts

The third line of effort is at the *operational* level, within the EU Military Staff. It focuses on the overall objective of defence modernisation through the *digitalisation of armed forces*. The latter is a multi-year defence modernisation effort towards network-centric warfare and network-enabled operations (a process Europeans started in the late 1990s) in which AI is a strategic enabler.[57] In this context, AI's highest likely impact in Intelligence, Surveillance and Reconnaissance (ISR) and cyber defence applications.[58] The added value of AI resides in enhanced situational awareness through multi-sensor information fusion and processing and decision-making assistance, which explains the European increased investment in sensor technologies. However, the AI use cases are not grounded in new operational concepts and doctrine but are treated as optimising accessories and layers to existing military capabilities.

In addition, AI enables European armed forces to improve efficiency across three dimensions: to *do things better*, notably, to increase their military effectiveness, readiness, and manoeuvrability, reduce resources and operational footprint; to *do better things*, such as improving mission planning, execution, and decision-making; and to *do new things*, such as sharing a common operating picture across the strategic, operational, and tactical levels, at national and multinational (EU) levels or sharing real-time information and coordinating across the civilian-military spectrum.[59]

There is a disconnect between the EU's political and military levels which obstructs a broader AI impact on strategic autonomy by putting a normative glass ceiling on innovation. The political level holds a deeply normative view of AI adoption in defence, with a desire to limit uses of autonomy and control technology diffusion. The technical and operational levels are deprived of any clear steer on adopting AI technologies to achieve a European military technological and operational advantage and of sustained political support for their plans. Meanwhile, there is no structured EU consideration of the impact of its normative and over-regulatory approach to AI on future European military advantage. Absent a convergent political and military approach to the integration of AI technologies in defence for the purpose of creating military advantage for the EU, AI will remain a distraction as far as EU strategic autonomy is concerned because algorithmic military power will remain outside the reach of the Union. In the next section, I identify and explore four reasons which explain this situation.

AI for autonomy: can AI contribute to greater strategic autonomy in defence?

Digital blueprints for technological autonomy

Whether one subscribes to the narrow or the wider understanding of strategic autonomy, *technology* is at the core of the concept. European Council conclusions,[60] the EU Global Strategy,[61] the Implementation Plan on Security and Defence,[62] and the EDF regulation[63] refer to strategic autonomy in the context of a sustainable, innovative, and competitive European defence technological and industrial base (EDTIB). The EDF specifically refers to the EDA's Capability Development Plan and the Overarching Strategic Research Agenda, which define current and future capability and technological requirements, as guiding instruments in the selection and implementation of EU-funded defence projects. The same is true for the EDA's AI in Defence Action Plan and Strategic Research Agenda, its work on Cyber Defence and its upcoming EDTs Action Plan.

Technological autonomy in defence is closely linked to European defence industrial competitiveness and it has been at the core of strategic autonomy since 2013.[64] It refers to the EU's ability to fund, develop, adopt, and integrate in defence capabilities critical technologies that are primarily (or exclusively) developed in Europe, are controlled by Europeans, and have secure supply chains. It is reflective of strategic European dependencies on raw materials and rare earths for the development of their military capabilities[65] as well as on foreign technologies that are subject to external export controls.[66]

This marks an evolution in the way the EU conceptualises the relationship between AI, cyber power, and strategic autonomy in defence, which was only implicit prior to 2019. At the 2018 EDA Annual Conference, former HR/VP Mogherini highlighted AI "is also a matter of security."[67] The May 2019 *Food for thought paper on Digitalisation and AI in Defence* acknowledged the role of AI as enabler of greater digitalisation[68] without specifically linking it to the objective of strategic autonomy.

By 2021, the EU's strategic thinking about the role of cyber and AI (as well as other digital EDTs) as indispensable elements of power, that shape the global balance of power, had evolved. The EU's 2021 review of the *Cyber Policy Framework* and the *Military Vision and Strategy on Cyberspace as a Domain of Operation* prioritise the full integration of cyber resilience, cybersecurity, and cyber defence "into the wider area of security and defence"[69] alongside other emerging technologies such as AI.

Furthermore, the EU Military Committee has developed a *Strategic Implementation Plan for the Digitalisation of EU Forces* which sets a level of ambition, specific targets, and milestones for the digitalisation and interoperability of European armed forced and discusses the role of AI as an enabler of digitalisation in defence. Finally, the Commission, the EDA and individual member states are trying to adapt their industry engagement strategies to enhance their access to cutting-edge technologies and accelerate adoption. In short,

AI is very much at the centre of European technological autonomy. The development of European AI technologies for defence can contribute to less European dependence on external actors for critical technologies associated with their military power.

However, EU (and individual national) efforts are not the only determinants of strategic autonomy. Exogenous factors are increasingly important in relation to EDTs, as highlighted by the proliferation of the use of concepts like "the geopolitics of technology." Among such exogenous factors is the relationship with private industry which is largely responsible for the rapid progress and accelerated investment in both AI and cyber. Continued European dependency on digital technologies and infrastructure (e.g., military cloud) from non-EU countries and industry actors entails the Union's freedom of action is inherently linked to its ability to build consensus with these external actors or risk tense geopolitical relations.

The structure and operation of the global internet may also act to inhibit European strategic autonomy independent of or in combination with other EU digital strategic dependencies. A normative approach to cyberspace and AI integration in defence without the backing of a technological foundation, is feasible only so long as the technological decoupling between the US and China does not accelerate and the internet is split into American and Chinese digital sphere of influence, which some analysts believe it is already the case (i.e., splinternets). Geopolitical alignment patterns are also important influence networks. For example, geopolitical constructs like the Belt and Road Initiative (BRI) and emerging strategic cooperation between the US and Indo-Pacific allies increasingly act as 'technological transmission belts' and 'techno-spheres of influence' with geopolitical impact.[70]

Finally, while the EU's regulatory power can shape market rules for AI technologies and it has established processes to negotiate rules for the "Responsible use of military AI," the Union cannot similarly shape global rules of operational deployment for AI and cyber in defence by itself. As EU external action takes place predominantly in a multilateral context, this inherently means the EU will require a new perspective on tech-diplomacy and partnerships,[71] which goes beyond the economic and trade field and addresses key defence and industrial issues related to competitiveness. However, building effective and operational partnerships has not been a priority for the EU. And it remains unclear how much the Union's approach will change in form and substance under the Strategic Compass. If the EU's security and technology partnerships will be driven by the same normative logic of the cybersecurity and hybrid threats toolboxes rather than a geopolitical one, they will arguably not lead to any significant enhancement of the EU's strategic autonomy in defence.

Digital blueprints for decision-making autonomy

Decision-making autonomy refers to the ability of the EU and European states to define their own strategic interests, collect and analyse data, and make

decisions independently of other international actors. The adoption of AI and the switch to "data-driven" policymaking requires further consideration of two aspects: what is the role of AI technologies in how EU decision-making in security and defence is taken and what are the AI technologies precisely used for in this process. Neither of these aspects is evident in the EU's approach to AI adoption in defence, which serves to underline the strategic difference between developing AI technologies and effectively employing them as instruments of statecraft.

It remains unclear how and for what purposes the EU can use AI technologies in decision-making to enhance its strategic autonomy. Based on the EU's AI initiatives in defence, the priority is the automated analysis of growing volumes of data in (near) real time. The May 2019 EUMS paper on AI and the digitalisation of the armed forces references the use of AI to gather, fuse and process large volumes of data at ever greater speeds for the generic purpose of "information superiority." According to its webpage, the European Space Agency (ESA) is already using algorithms to analyse satellite imagery, though not necessarily in defence-related scenarios. The EDA's *Action Plan on AI in Defence* resulted in three priority use cases: AI for cyber defence, AI for ISR and enhanced situational awareness and AI for smart maintenance.[72] The EU's *2021 Cybersecurity Strategy* refers to the use of AI to monitor and manage digital networks and develops the EU's institutional structures that support cyber security, including through the creation of a Joint Cyber Unit, an EU cyber intelligence working group and a network of AI-enabled cyber centres across the member states. Beyond this, EU can better leverage AI for decision-making support and CSDP mission planning and execution. For example, the EEAS is exploring the notion of using AI-enabled dashboards for conflict prevention and early warning[73] while the EUMS is interested in AI-enabled mission planning for different scenarios.

At the political level, strategic autonomy has been linked with discussions about the reform of EU decision-making on security and defence. Specifically, this refers to the proposal to adopt QMV in the Council and activate article 44 TEU on issues pertaining to foreign policy as a means of facilitating swifter, more agile and more unified EU positions and responses to international developments.[74] Notwithstanding academic debates about the virtues of efficiency and legitimacy in EU security and defence decision-making, QMV and art 44 TEU are geared towards political efficiency by bypassing the current consensus rule. This is in recognition of the fact that in responding to crises, speed is a critical consideration whereas building consensus is often a time-consuming process that frustrates rapid response and leads to the minimal common denominator among the EU members on the type and scope of response. However, it's unclear how AI can be leveraged alongside QMV and art 44 to deliver a more efficient EU "data-driven" decision-making process in foreign, security and defence policy. Indeed, integrating AI and AI-generated information into European political and military decision-making

may pose several challenges to the QMV model which need to be accounted for. Three categories of challenges are telling.

First are *procedural* challenges, such as the origin and reliability of the data and algorithms used in the decision-making process and national and institutional resistance to data sharing. The adoption of AI also adds a layer of complexity to military and civil-military decision-making by requiring a reconsideration of appropriate levels of decision-making for specific AI-supported actions and dedicated efforts to foster a shared civil-military operating picture and threat understanding.

Second, *normative* challenges should not be underestimated. These include whether the actionable information generated by AI systems is compliant with European values and it is trusted by decision-makers from different nations, with different access to comparable levels of intelligence and technology. While it is true for NATO that "the sense of equality and co-decision among members could be at risk because of worries about accountability,"[75] this also applies to EU decision-making. Whether in cyberspace or in the physical domains of warfare, the adoption of AI implicitly creates requirements for greater use of autonomy which conflict with different normative policies of many EU member countries.

Third, there are important *foundational* challenges to consider. Some of these emanate from the mismatch between technological potential and the specific problem(s) AI technologies are meant to address. AI algorithms are modelled on optimising performance in specific contexts and instrumentally dealing with cognitive tasks. They are not built to support consensus and a common understanding which are the main challenges for EU decision-making. Arguably, AI-enabled decision-making assistance could challenge pluralistic decision-making among various European actors, with different interests and ideological preferences. For example, AI can reveal patterns of adversarial Russian and Chinese behaviour. However, it will not eliminate diverging national preferences or lead to a coherent EU policy on either of these rival great powers.

Other foundational challenges emanate from the *nature of the action* and *threat* that is subject to decision-making. The EU's decision-making model – regardless of whether it is QMV, art 44 or another – remains essentially reactive. The use of AI-enabled capabilities and enhanced situational awareness could make a predictive and/or pre-emptive model possible and it could even automate some defence responses. However, the EU is unprepared to fully leverage this technological potential without a radical shift in its strategic culture.

Managing expectations about what type of European action AI would enable is important in understanding whether and how (much) AI technologies support European strategic autonomy. Specifically, an important limitation for AI technologies in the context of EU decision-making is that it cannot generate or replace political will to act. In short, AI technologies will not cure Europe's 'bias for inaction.' The risk of reductionist techno-centric

approaches cannot be discounted and absent an EU common approach to AI in defence, these challenges will not be easy to bypass regardless of the decision-making format.

Digital blueprints for operational autonomy

Towards an EU digital military level of ambition?

Since 2016, the Union sought to address some of its main challenges through PESCO, the EDF, the CARD, EU-NATO cooperation and the *Strategic Compass*. A well-equipped and operational rapid reaction force, the consolidation and competitiveness of its defence industrial base, the development of modern military capabilities and the development of an autonomous military command and control and planning tool remain critical to (self-) perceptions of European strategic autonomy. In other words, operational autonomy and military capabilities are key to European strategic autonomy.

Operational autonomy refers to the ability of European states to jointly develop competitive modern military capabilities, in sufficient numbers, levels of interoperability and readiness and to be able to successfully launch, conduct, and sustain military operations across the full spectrum from low-to-high intensity warfare independent of external (i.e., American / NATO) support.

The creation of an EU operational and well-equipped rapid reaction force (rather than *ad hoc* multilateral coalitions) is regarded as a "clearly decisive step towards European defence."[76] However, the EU's level of military ambition does not match its policy rhetoric. In the context of the Strategic Compass, several EU member states agreed to launch a joint "first entry" force of 5000 troops, possibly based on the activation of the EU Battlegroups.[77] In addition, to enhance its operational autonomy, the EU is also developing an EU Full Spectrum Force Package, an EU Strategic Reserve Concept, an EU CSDP strategic stockpiling for military CSDP Operations and Missions as well as an EU Concept for Military Command and Control. Nevertheless, to put things in strategic perspective, the EU's 5000-troops rapid response force (essentially, the EU's version of VJTF), which is expected to undertake two or more simultaneous operations of various scope and intensity, is smaller than the American contingent securing Kabul airport during the August 2021 withdrawal and is roughly similar in size to the estimated size of Russian troops operationally deployed in Syria. Furthermore, it is unclear how and whether the EU will overcome significant force generation, common funding, and decision-making challenges in implementing the first entry force.[78]

The scope of the EU's military level of ambition – and therefore, the scope of its operational autonomy – are linked to the ongoing process in the Strategic Compass to

> reflect on realistic contingencies in light of the ring of instability and tension around Europe and beyond. In essence, *in addition to preparing*

> *for one major contingency* (which has subsequently been complemented by several other illustrative scenarios), the European Union now needs to focus much more on *preparing to undertake simultaneously several smaller and medium-sized operations – and not only on land, but also at sea and in the air.*[79]

This is not new – it essentially describes the EU's military level of ambition under the Headline Goals which have yet to be achieved. The caveat, of course, is that under the Headline Goals, EU operations were not necessarily based on scenarios that required either long-distance power projection or the EU's ability to operate in highly contested and denied environments like the Indo-Pacific, Eastern Europe, or the High North.

In addition to national efforts on integration of AI in defence, the EU Military Committee has developed a *Strategic Implementation Plan for the Digitalisation of EU Forces* which sets a level of ambition for the digitalisation of European armed forces, including the development of a full spectrum, scalable digitalised force package,[80] and provides an implementation roadmap. According to the EU Military Staff's definition, *digitalisation of defence* refers to the application of multiple information technologies (including AI) to acquire, process, disseminate, and use information across the multi-domain battlespace, by networking sensors, capabilities and forces and decision-makers and enabling the achievement of information superiority.[81] Therefore, the concept of *digitalisation of defence* highlights the interdependence between AI and cyber power.

Digital futures: AI and cyberwarfare

It is worthwhile briefly exploring how technological convergence between cyber and AI technologies is relevant for modern warfare and European strategic autonomy.

AI technologies used in military applications have the potential to revolutionise the way wars are fought. In cyberspace, the frequency of adversarial interaction, the variety of actors and the challenge of attribution translates into the added value of increasing automation of cyber defensive and offensive tools. Cyber is not just a new domain of operations, but cyber power is "a new form of power projection" for great powers that exploit both the military and civilian capabilities, notably "sometimes employing companies under their jurisdiction and control as their agents."[82] These perceptions are grounded in the higher frequency and intensity of cyber-attacks during the COVID-19 pandemic, including but not limited to the Solaris Winds Orion attacks. They are also grounded in the strategic great power competition which is altering the structure of the internet and putting a premium on securing data as a strategic asset. Countries such as Russia, China, Iran, and North Korea control the digital sovereign space, notably cyber capabilities, the national internet, digital infrastructure, and data flowing through them.

The relationship between AI technologies and cyberspace is one of interdependence. As a digital category of technologies, cyberspace is the primary domain for AI. As such, this paper considers both the role of AI technologies in cyberspace for cyber resilience, cyber defence, and cyber offense capabilities as well as the role of cyber security and cyber defence for the viability of deployed AI military applications.

First, AI technologies contribute to the automation of cyber offensive tools.[83] The relevance of AI-enabled cyber offense stems from "new motivations for operations within the cyber domain" which "differ dramatically from the more conventional digital threats,"[84] creating "a new class of persistent threat tools" and actors[85] and challenging current assumptions on cyber offense-defence balance and cyber conflict prevention. Such transformations are grounded in the greater sophistication, speed, reach (i.e., attack surface), adaptability, and complexity of AI-enabled cyber capabilities for both defensive and offensive purposes, all of which are beyond the human speed of reaction and capacity to control. AI-enabled cyber offensive tools could seek to illicitly access data and learn from it the vulnerabilities of target systems, or it can adapt throughout the mission to select from a range of options on how to proceed and infect the target network. However, "the prospect of subverting AI-driven security functions (…) incentivizes operations in cyberspace beyond in-domain effects and outcomes."[86] In other words, a cyber-attack could result in physical effects. Examples include cyber-attack on armed unmanned vehicles used to attack civilians or friendly troops.

Second, for cases in the virtual and physical domains, AI technologies hold the same promise of improving cyber resilience and cyber defence through the deployment of automated network monitoring and early detection and response agents. Though progress has been slower, AI algorithms can be trained to detect and automatically patch software vulnerabilities in human and digitally enabled code writing,[87] or they can detect, neutralise, and respond to a cyber-attack on physical military platforms as in Purdue University's AI2I project.

Third, AI systems are particularly vulnerable to cyber-attacks, either through "input attacks" which mislead the AI by skewing its pattern recognition through deceptive measures or through "poisoning attacks" which target and corrupt the code of AI algorithms and the data they use.[88] Examples of AI-enabled cyber defence systems are proliferating – Automatic Intelligent Cyber Sensor and Enterprise Immune System are just two such examples.

However, a note of caution is warranted. AI systems used in cyber are still faced with technological maturity challenges as well as with networking, infrastructure, and resource challenges[89] and their efficiency varies depending on the exact model used (deep or machine learning). Manned or unmanned complex military platforms such as submarines, fighters, drones, and armoured vehicles may require teams of autonomous cyber defence agents to be operationally deployed on the battlefield. Machine-to-machine interaction protocols and autonomous cyber agents teaming protocols, bandwidth,

electrical and computing power consumption are important challenges to the rapid adoption of operationally deployed AI-enabled cyber agents. Particularly challenging is that autonomous cyber agents may need to operate for longer periods of time with limited or no human in the loop when deployed in highly contested battlefields. As a result, their activity could cause the malfunction of military systems they are meant to protect,[90] and such malfunctions could take longer to discover absent close interaction with human operators.

In other words, while technology maturity levels for state-of-the-art AI may not enable their deployment and efficient operation in real-life operational conditions in the medium term, fully automated cyber capabilities are not just possible in the foreseeable future but may indeed become altogether necessary. As likely "primary cyber fighters on the future battlefield" stealthy, resilient, and multipurpose autonomous cyber agents will be critical for cyber defence as well as for the cyber security of other actors, including forces and military platforms on the battlefield.[91]

Outstanding questions

In the context of new operational requirements of algorithmic warfare, EU planning efforts in AI and cyber are a step in the right direction. However, there are several outstanding questions regarding the European adoption of AI in defence, including whether under the given circumstances the EU represents the appropriate format and level for implementing and conducting algorithmic warfare. Ongoing efforts fall short of a common European approach to emerging technologies or a military strategy which would fully integrate cyber and AI-enabled capabilities into CSDP multi-domain operations, identify capability targets and milestones, direct investment and defence planning, utilise autonomous European intelligence about AI and cyber threats posed by strategic rivals and clarify military and political decision-making procedures around the use of AI and autonomy in cyber, space and the physical domains of operations.

On the policy and planning front, it is unclear whether EU efforts to incentivise and support the development of AI-enabled capabilities will be successful across a wide range of European armed forces. Equally, it is unclear whether the current EU military level of ambition, doctrine and training are adequate for AI integration, for the requirements of cyber security and resilience of AI-enabled military applications, and issues related to industry and the security of digital supply chains. There is currently no adequate understanding of the required level of cyber security in European mission critical systems and platforms. On the technical side, it is unclear whether the EU member states can overcome their capabilities fragmentation which potentially puts European strategic autonomy at risk.[92]

Even relatively low-hanging fruit such as AI-enabled logistics and maintenance systems for multinational operations are challenging for European

states because they operate in a low-trust environment.[93] Sharing data into common 'data lakes' to facilitate the development of common capabilities and platforms used in CSDP missions and operations is still not possible and whatever national AI-enabled maintenance capabilities are currently under development would not be deployable for security reasons,[94] albeit enhanced space-based communications capabilities could change this in the future.[95] Significant concerns also exist about the lack of European deployable advanced C4ISR battle networks[96] which means the utility of individual AI-enabled capabilities is rapidly undermined by exceeding security risks.

Autonomy by partnership? AI and EU-NATO cooperation

The Implementation Plan on Security and Defence argues "Europe's strategic autonomy entails *the ability to act and cooperate with international and regional partners wherever possible*, while being able to operate autonomously when and where necessary."[97] Nevertheless, this poses a dilemma and a trade-off for the EU in developing *and* operating AI-enabled capabilities and cyber power autonomously or in cooperation with partners.

The EU-NATO cooperation and EU-US cooperation are cases in point. European strategic autonomy requires greater operational capability, including autonomous command and control. Nevertheless, in the context of the digitalisation of defence, some EU officials are keen to proceed with the full alignment of standard operating procedures with NATO's Framework Network and the Federated Mission Concepts which are currently blocked by Turkey.[98] The EU requirements for military mobility were established based on NATO minimal requirements for infrastructure. This points to great interdependence between the EU and NATO and cyber and EDTs are no exception.

EU and NATO representatives signed the 2016 and 2018 Joint EU-NATO Declarations which cover cyber, hybrid, and capability development as priority areas of cooperation. Cyber is one of the areas of EU-NATO cooperation with the most consistent reported progress over the past five years,[99] across both qualitative and quantitative assessment metrics. NATO's own efforts in the field of developing cyber doctrine, structures, capabilities, and a common strategic culture span almost two decades. Lessons learned from NATO's efforts to integrate cyber into its operations, protect its own critical networks and infrastructure and make decisions on deploying cyber effects are fully relevant to the EU's efforts. A case in point is the differentiated integration of defensive and offensive cyber effects in NATO operations.

As AI technologies become increasingly incorporated into cyber capabilities, tasks and operations, closer EU-NATO cooperation on AI uses in the cyber domain seems a natural next step. The foundations for such cooperation already exist. Both NATO and the EDA (and ENISA) have explored the role of AI and automation in cyberspace. In 2016, the NATO Science and Technology Organisation (STO) established a research group on *Intelligent*

Autonomous Agents for Cyber Defence and Resilience which proposed a reference NATO architecture for Autonomous Intelligent Cyber-Defence Agents based on their anticipated key functions and technical requirements as well as a roadmap towards their adoption.[100] EDA projects such as Cyber Defence Technology Landscaping project, CYSAP, CHESS, and GARD seek to support EU member states in using AI technologies for cyber threat mapping, enhanced automated network resilience and human-machine cooperation for early detection of cyber incidents. Similarly, AI for cyber defence is one of the first pilot projects under both NATO's 2021 AI Strategy and one of three priority actions under the EDA's Action Plan for AI in Defence. This is a substantial foundation for closer EU-NATO cooperation in AI-enabled cyber defence.

However, as the EU and NATO negotiate the third EU-NATO Joint Declaration which will reportedly focus on EDTs and climate change cooperation, the level of ambition for EU-NATO cooperation remains vague.[101] A 2021 German-Dutch Food for through paper advocated a new [EU-NATO] "joint declaration" establishing "stronger political consultations" of a joint informal working group on EDTs.[102]

Finally, closer cooperation also entails a deliberate effort on the part of the EU and NATO to avoid competition.[103] Both EU and NATO defence innovation landscapes are becoming more complex, with new bodies and instruments recently established. In May 2021, the European Council called for "reinforcing the role played by the EDA in fostering defence innovation including disruptive technologies"[104] and in June 2021 NATO leaders agreed to establish a Defence Innovation Accelerator for the North Atlantic (DIANA), backed by a dedicated NATO Innovation Fund (NIF) reportedly worth $1bn. Both the EU and NATO's new innovation structures are experiencing early challenges – the US and France, two of the lead innovators in NATO are reluctant to join DIANA and NIF, whereas the push for the EDA's greater role in defence innovation is almost exclusively backed by Paris. Fears of duplication (especially on the European side) are pervasive – though there is little evidence of ongoing efforts to ensure complementary by design between these complex innovation structures and even less evidence of why duplication in research and innovation is necessarily problematic. The current focus in the EU and NATO frameworks seems to be on form rather than substance with the (still unproven) expectation that once appropriate innovation structures are in place, substance – and strategic results – will follow. Along with the challenges in transatlantic defence cooperation on AI and other emerging technologies[105] this is a reminder that (the degree of) European strategic autonomy may not necessarily be a choice, but a necessity.

Conclusion

Europeans are arguably still in the incipient phases of their AI efforts and the bulk of the difficult tasks of developing and widely adopting AI technologies

still lie ahead. The EU is off to a good start in planning its technology policies and seeking cross-fertilisation as a long-term strategy of technology competition with and strategic autonomy from other great powers.

While AI and cyber convergence has significant strategic implications for European security, this paper proved that a techno-centric European strategic autonomy is a distraction. AI adoption in defence will add a new layer of complexity to a complicated multinational and institutional landscape, without a clear perspective on how it will be translated into an effective tool of statecraft for greater strategic autonomy. This finding is substantiated by empirical evidence across four dimensions of strategic autonomy: technological, operational, decision-making autonomy and defence partnerships.

This paper has barely scratched the surface of the research needed on the impact of AI on European strategic autonomy in defence. It was not meant as an exhaustive ontological analysis of AI and European strategic autonomy or as a dissuasive narrative towards greater EU efforts to integrate AI, cyber effects into defence. Rather it was intended as a pragmatic analysis of four critical dimensions of strategic autonomy and AI's impact on them in relation to cyberspace and other domains of warfare. More research is needed to understand how or whether EU planning translates into qualitatively different AI-enabled capabilities, how differentiated AI adoption among the EU states influences their collective capacity for common action and how the EU employs its 'algorithmic power' across strategic partnerships, networks, and channels.

The research findings of this paper have theoretical and conceptual implications beyond the European context. Its findings on the conceptualisation of the empirical relationship between technology and political autonomy for action is potentially relevant across a wide range of international organisations, alliances, and multilateral collaborative fora of different levels of institutionalisation. It has direct bearing on understanding the role such multilateral actors enabled by technology can play in great power competition. It also serves as a reminder of the perils of reductionist techno-centric approaches to strategic autonomy which underestimate the complex challenges of translating technological and innovation potential into actual tools of power in international politics.

Acknowledgement

The author is grateful to Dr Bastian Giegerich and Dr Joe Burton for their constructive comments on earlier drafts of this paper.

Notes

1 Such technologies include artificial intelligence (AI), quantum computing (QC), big data analytics, more resilient cyber networks, cloud and edge computing, sensor technologies, next generation (tele)communications, space technologies, chip and semiconductor technologies, and their supply chains and functional integration.

2 Ministère des Armées, "Strategic update 2021," (February 2021): 26, 38, 41.
3 German presidency of the European Council, "Independent, inclusive and innovative: Four goals of the German Presidency for the digital sector" (2020).
4 David M. Herszenhorn, "Biden's top security adviser sees strong transatlantic alliance (and no jumping in lakes)," *Politico*, 8 October 2021; Matthew Burrows and Julian Mueller-Kaler, "Europe's third way," *Atlantic Council*, 14 March 2020; Josep Borrell Fontelles, "The Sinatra doctrine: Building a united European front," *Institute Montagne*, 9 September 2020; and Hal Brands, "Germany is a flashpoint in the U.S.-China cold war," *Bloomberg*, 23 February 2021.
5 Andrey Kortunov, "Russian perspective on the challenges to the European project," *Russian International Affairs Council*, 9 July 2021.
6 For example, President Charles Michel recently argued: "We are sending a message not only to our citizens, but also to the rest of the world: Europe is a world power. We are ready to firmly defend our interests." Council of the European Union, "Recovery plan: Powering Europe's strategic autonomy – Speech by President Charles Michel at the Brussels Economic Forum," 8 September 2020.
7 Carla Hobbs (ed.), "Europe's digital sovereignty: From rulemaker to superpower in the age of US-China rivalry," *ECFR*, 30 July 2020; Ulrike Franke and Jose Ignacio Torreblanca, "Geo-tech politics: Why technology shapes European power," *ECFR*, 15 July 2021.
8 European External Action Service, "Shared vision, common action: A stronger Europe. A global strategy for the European union's foreign and security policy" (June 2016): 4, 9.
9 European Council, "Special meeting of the European Council (1 and 2 October 2020): Conclusions," *EUCO 13/20*, 2 October 2020: 1.
10 Council of the European Union, "Foreign affairs council conclusions on implementing the EU global strategy in the area of security and defence" (14 November 2016): 1. This is notwithstanding the different views held by European states, the lack of a common understanding of this foundational concept among the EU states and the proliferation of seemingly alternative or complementary concepts, such as "strategic sovereignty," "strategic responsibility," "strategic resilience" and "open strategic autonomy." These aspects are beyond the scope of this paper, but for a good analysis see Hans-Peter Bartels, Anna Maria Kellner, and Uwe Optenhögel (eds.), *Strategic Autonomy and the Defence of Europe: On the Road to a European Army?* (Berlin: Dietz, May 2017).
11 Operational autonomy inherently entails the development of relevant military capabilities, too. However, this aspect is well documented in international relations literature by comparison to the operational dimension of strategic autonomy, discussed in this paper.
12 Josep Borrell Fontelles, "Why European strategic autonomy matters," 3 December 2020; and Council of the European Union, "Council conclusions on security and defence, 8396/21," 10 May 2021: 2.
13 Notably, the intensification of renewed great power competition, deceleration of globalisation, rapid technological progress and accelerated diffusion of technology, new domains of coercive power (cyber and space), and the changes within the transatlantic partnerships after the election of Joe Biden.
14 For example, President Charles Michel argued on two occasions that "European strategic autonomy is goal No. 1 for our generation" and "The strategic independence of Europe is our new common project for this century." See Council of the European Union, "Strategic autonomy for Europe - the aim of our generation - speech by President Charles Michel to the Bruegel think tank," 28 September 2020; and Council of the European Union, "Recovery plan."
15 Council of the European Union, "Digital sovereignty is central to European strategic autonomy - Speech by President Charles Michel at "Masters of digital 2021" online event," Press release, 3 February 2021.

16 Barbara Lippert et al., "European strategic autonomy: actors, issues, conflicts of interests," *SWP Research Paper* (March 2019): 5; and Giovanni Grevi, "Fostering Europe's strategic autonomy – A question of purpose and action," *EPC* (December 2020).
17 Council of the European Union, "Council conclusions, 8396/21," 7–8.
18 Strategic sovereignty "signifies the ability to act autonomously, to rely on one's own resources in key strategic areas and to cooperate with partners whenever needed. To fully develop such strategic sovereignty, the EU needs to show political will and strengthen its capacity to act." Suzana Anghel, Beatrix Immenkamp, Elena Lazarou, Jerôme Leon Saulnier, and Alex Benjamin Wilson, "On the path to 'strategic autonomy': The EU in an evolving geopolitical environment" (September 2020): 1.
19 Within the limits of the legal authorities established under the EU treaties.
20 This aspect is implied by the focus, within the context of the strategic compass, on flexibility between CSDP and ad hoc missions, operations, and defence initiatives of the member states. In addition, PESCO and EDF projects support the development of capabilities that are fully owned and operated by the participating member states, which under the principle of "single set of forces" can be used in EU missions and operations as well as in other multilateral formats, such as NATO, the UN and other European minilateralism defence initiatives. Moreover, since EU states differ significantly in their capacity to adopt AI in the medium and long term, AI technologies could conceivably empower ad hoc European coalitions to deal with small or medium scale contingencies of lower intensity, within or outside the EU framework. This would also be consistent with growing Franco-German desire to preserve their national freedom of manoeuvre and decision-making liberty on all matters in security and defence.
21 Josep Borrell Fontelles, "What's next for European defence?" *European External Action Service*, blog post, 7 May 2021.
22 Vincenzo Camporini et al., "European preference, strategic autonomy, and European defence fund," *IRIS ARES* (November 2017); Hugo Meijer and Stephen G. Brooks, "Illusions of autonomy: Why Europe cannot provide for its security if the United States pulls back," *International Security* 45, no. 4 (2021): 10.
23 Douglas Barrie, Ben Barry, Henry Boyd, Marie-Louise Chagnaud, Nick Childs, Bastian Giegerich, Christian Mölling, and Torben Schütz, "Protecting Europe: Meeting the EU's military level of ambition in the context of Brexit," *IISS Research Paper* (29 November 2018).
24 David Bachmann, Tobias Bunde, Quirin Maderspacher, Adrian Oroz, Gundbert Scherf, and Kai Wittek, "European defense report. More European, more connected, more capable: Building the European armed forces of the future" (2017); and Max Bergmann, James Lamond, and Siena Cicarelli, "The case for EU defense: A new way forward for trans-atlantic security relations," *Center for American Progress* (1 June 2021).
25 Dick Zandee, Bob Deen, Kimberley Kruijver, and Adája Stoetmans, "European strategic autonomy in security and defence: Now the going gets tough, it's time to get going," *Clingendael Report* (December 2020).
26 The exact EU military level of ambition is in flux. A 2016 preparatory conference for the EU Global Strategy, organised by the Netherlands, discussed an EU military level of ambition comprising 10,000–15,000 troops high-readiness force – well below the 1999 Helsinki Headline Goal. In the context of the strategic compass, the idea resurfaced in 2021 in the form of a "first entry force" comprising roughly 5000 troops established by 14 EU member states and based on the roughly 1500-troops EU Battlegroups as its core. Borrell, "What's next for European Defence"; and Robin Emmott and Sabine Siebold, "EU should

enable military coalitions to tackle crises, Germany says," *Reuters,* 2 September 2021. The EU's level of military ambition in relation to crisis management is still under debate as part of the scenario-based approach of the Strategic Compass. See: Arnout Molenaar, "Unlocking European defence. In search of the long overdue paradigm shift," *IAI* (22 January 2021): 9. In the aftermath of the messy Afghanistan withdrawal, HR/VP Borrell even referenced the need for a 50,000 troop EU rapid reaction force, which would mean returning to initial corps level established under the Headline Goals.

27 According to the *Joint Research Centre*, no European chip manufacturer offers products with features over 22 nm whereas leaders in this offer products with 5 nm (Taiwan, South Korea) and 7 nm (the US) features which are rapidly becoming critical for advanced computing techniques, including AI and QC.

28 Kincsö Izsak, Maialen Perez, Henning Kroll, and Sven Wydra, "Advanced technologies for industry – EU report technological trends and policies," *Joint Research Center* (November 2020): 50–55, 82.

29 Roland Berger, "The road to AI – Investment dynamics in the European ecosystem. AI global index 2019," *France Digitale* (1 January 2019): 8.

30 Claudiu Pavel and Evangelos Tzimas, "Raw materials in the European defence industry," *Joint Research Center* (2016): 6–11.

31 Claudiu Pavel and J. Huisman, "Critical raw materials for strategic technologies and sectors in the EU: A foresight study," *Joint Research Center* (14 September 2020).

32 Notably, to encourage widespread adoption of AI technologies across the EU with a view to maximising competitiveness, to prepare for socio-economic changes determined by the adoption of AI and to develop adequate ethical and legal frameworks to support AI adoption.

33 It has done so through four sets of concrete actions: accelerating AI development and uptake across the EU, enabling AI innovation to cross the valley of death from the lab to the marketplace, developing the right policy framework for trustworthy AI and building European leadership in high-impact sectors of AI, including robotics and law enforcement. European Commission, "Coordinated plan on artificial intelligence 2021 review," 21 April 2021: 2–4.

34 Spin-offs refer to the use and adoption of emerging and disruptive technologies developed through defence funding and channels into the broader society and economy, whereas spin-ins refer to the successful adoption and integration of civilian-developed technologies into defence applications.

35 European Commission, "Europe's digital decade: Digital targets for 2030," 9 March 2021.

36 European Commission, "Action plan on synergies between civil, defence and space industries," *COM (2021) 70 Final*, 22 February 2021: 10–11, 15–16.

37 European Commission, "Joint communication to the European parliament and the council: The EU's cybersecurity strategy for the digital decade," *JOIN (2020) 18 Final*, 16 December 2020: 5–6, 17–18.

38 European Union Agency for Cybersecurity (ENISA), "AI cybersecurity challenges: Threat landscape for artificial intelligence" (December 2021): 31.

39 Simon Pfeiffer and Randolf Carr, "Trust not found: A European survey on digital (dis)trust," *Munich Security Conference Brief*, no. 2 (March 2021): 3.

40 Munich Security Conference, "Munich security index 2021: Appendix to the Munich security report 2021. With Additional Survey Results and Analysis" (February 2021): 26.

41 Simona R. Soare, "European military AI: Why regional approaches are lagging behind," in *Global Strategic Perspectives on Military AI* (Singapore: Routledge, 2022).

42 Izsak et al., "Advanced technologies," 9.
43 In 2020 the Commission and the European investment fund launched "the first six Venture Capital funds under the InnovFin Artificial Intelligence and Blockchain pilot" with a cumulative budget of €700 million to support the adoption of AI and blockchain technologies. European Commission, "First six Artificial Intelligence and blockchain technology funds backed by innovfin raise a total of EUR 700 m," 28 October 2020.
44 These refer to the central and emphatic role of European values, human rights, market principles, regulatory standards and criteria across all relevant documents studied in this paper.
45 European Defence Agency, "R&T conference – Impact of disruptive technologies on defence. Speech by Jiří ŠEDIVÝ, EDA Chief Executive," 20 April 2021: 3.
46 This comprises activities deriving from the specific priorities of individual presidencies, ongoing work on AI across several preparatory bodies and working parties within the Council and the Parliament's Foreign Affairs Committee, Subcommittee on Security and Defence and Special Committee on Artificial Intelligence.
47 Simona R. Soare and Fabrice Pothier, "Leading edge: Key drivers of defence innovation and the future of operational advantage," *IISS Research Paper* (November 2021).
48 Ulrike Franke, "Artificial intelligence diplomacy: Artificial intelligence governance as a new European union external policy tool," *European Parliament*, Directorate-General for Internal Policies, PE 662.926 (June 2021): 21–28.
49 European Parliament special committee on artificial intelligence in the digital age, "AIDA Working Paper on 'The External Policy Dimensions of AI' following the AIDA/AFET/SEDE public hearing on 1 and 4 March 2021," March 2021.
50 European Parliament, "European Parliament resolution of 12 September 2018 on autonomous weapon systems (2018/2752(RSP))," 12 September 2018; and European Parliament, "Artificial intelligence: questions of interpretation and application of international law in so far as the EU is affected in the areas of civil and military uses and of state authority outside the scope of criminal justice," 2020/2013(INI), 20 January 2021.
51 Notably, the Preparatory Action on Defence Research (PARD), European Defence Industrial Development Programme (EDIDP), European Defence Fund (EDF) and PESCO projects.
52 These aspects are naturally covered at the national level, though this creates interoperability and coordination challenges within CSDP as member states approach the use of AI in military applications very differently.
53 European Commission, "European Defence Fund – Calls 2021," *Factsheet,* 30 June 2021.
54 European External Action Service, "Questions and answers: a background for the Strategic Compass," Memo, 3 September 2021: 4.
55 Author interview with European Commission officials, November–December 2021.
56 European Defence Agency, "R&T Conference," 2–3.
57 Author interview with EUMC representative, November 2019.
58 Ibid.
59 Ibid.
60 European Council, "European council conclusions, EUCO 217/13," 19–20 December 2013: 7.
61 European External Action Service, "Shared vision," 46.
62 Council of the European Union, "Implementation plan on security and defence," 14392/16, 14 November 2016: 7.

63 Official Journal of the European Union, "Regulation (EU) 2021/697 of the European parliament and of the council of 29 April 2021 establishing the European Defence Fund and repealing Regulation (EU) 2018/1092," 29 April 2021: 2, 13.
64 European Council, "Conclusions, EUCO 217/13," 8.
65 Pavel and Tzimas, "Raw materials," 21–24; and Pavel and Huisman, "Critical raw materials," 70.
66 Valerio Briani et al., "The Development of a European Defence technological and industrial Base (EDTIB)," European Parliament, Directorate-General for External Policies of the Union Study, June 2013: 51–58.
67 European Defence Agency, "Federica Mogherini opens Annual Conference devoted to unmanned/autonomous systems," 29 November 2018.
68 Finnish Presidency of the Council, "Food for thought paper by Finland, Estonia, France, Germany, and the Netherlands: Digitalization and artificial intelligence in defence," 17 May 2019.
69 Council of the European Union, "Council conclusions, 8396/21," 15; and European Parliament, "Artificial intelligence," 7.
70 Simona R. Soare, "Politics in the machine: The political context of emerging technologies, national security, and great power competition," in *Emerging Technologies and International Security: Machines, the State, and War* (London: Routledge, 2020), 109.
71 Simona R. Soare, "European defence and AI: Game-changer or gradual change?" *RSIS Commentary* (24 March 2021): 7.
72 Author interview with EDA officials, December 2021.
73 Katariina Mustasilta, "Preventing our way back to friendship? Conflict prevention and the future of transatlantic relations," in *Turning the tide: How to rescue transatlantic relations* (EUISS, 2020), 109–110.
74 Anghel, "On the path."
75 Tomas Valasek, "How artificial intelligence could disrupt alliances," *Carnegie Europe*, 31 August 2017.
76 Ronan Le Gleut and Hélène Conway-Mouret, "European defence: The challenge of strategic autonomy," French Senate report no. 626 prepared for the Extraordinary Session of 2018–2019, 3 July 2019: 85.
77 Borrell, "What's next for European defence."
78 Brooke Tigner, "EU's proposed rapid reaction entry force faces many hurdles," *Jane's*, 11 May 2021.
79 Molenaar, "Unlocking European defence," 9.
80 Author interview with EUMC representative, November 2019.
81 Ibid.
82 L. Ilves and A-M. Osula, "The technological sovereignty dilemma – and how new technology can offer a way out," *European Cybersecurity Journal* 6, no. 1 (2020): 6.
83 Ryan Ko, "Cyber autonomy: Automating the hacker – self-healing, self-adaptative, automatic cyber defence systems and their impact on society, industry, and national security," in *Emerging Technologies and International Security: Machines, the State, and War* (London: Routledge, 2020), 177–180.
84 Christopher Whyte, "Poison, persistence, and cascade effects: AI and cyber conflict," *Strategic Studies Quarterly* 14, no. 4 (2020): 19.
85 Christopher Whyte, "Scenario 2—AI and insecurity for all: The future of cyber conflict," in *Alternate Cybersecurity Futures*, John Watts et al., Atlantic Council, September 2019: 12.
86 Whyte, "Poison, Persistence," 29.
87 Kim Martineau, "Deep-learning models code more like humans," *Control Engineering,* 23 April 2021.
88 Whyte, "Poison, persistence," 26–27.

89 Iqbal H. Sarker, Md Hasan Furhad, and Raza Nowrozy, "AI-driven cybersecurity: An overview, security intelligence modelling and research directions," *SN Computer Science* 2 (2021): 172–173; Robert Thomson, Christian Lebiere, and Drew Cranford, "Achieving active cybersecurity through agent-based cognitive models for detection and defense," *United States Military Academy of West Point* (2021); Yirui Wu, Dabao Wei, and Jun Feng, "Network attacks detection methods based on deep learning techniques: A survey," *Hindawi Security and Communication Networks* (2020): 14.
90 Alexander Kott et al., "Autonomous intelligent cyber-defense agent (AICA) reference architecture release 2.0," *US Army Research Laboratory*, September 2019: 62–64.
91 Ibid., 57, 61.
92 European Defence Agency, "2020 CARD report: Executive summary," 21 November 2020: 2.
93 Jan Techau, "Why the EU can't do security and defence," *EUObserver,* 23 October 2019.
94 Author interviews with EDA representatives, November–December 2019.
95 Jean-Marc Tanguy, "France launches first Syracuse IV telecommunications satellite," *Jane's,* 26 October 2021.
96 Author interview with EUMC representatives, November 2019.
97 Council of the European Union, "Implementation plan, 14392/16," 4.
98 Author interview with EUMC representative, November 2019.
99 This progress includes exchanges on concepts and doctrine, threat indicators, cyber capability development, mutual participation of EU and NATO staff in each other's education programmes, trainings, and exercises, the integration of cyber effects into crisis management and updates in cyber crisis management and response mechanisms. See NATO, "Progress report on the implementation of the common set of proposals endorsed by EU and NATO Councils on 6 December 2016 and 5 December 2017" 2016–2021.
100 Kott et al., "Autonomous intelligent."
101 See, for example, the vague language in Council of the European Union, "Council Conclusions, 8396/21," 7.
102 David M. Herszenhorn, "German, Dutch diplomats urge stronger NATO-EU ties," *Politico,* 20 May 2021.
103 Simona R. Soare, "Partners in need or partners in deed? How EU-NATO cooperation shapes transatlantic relations," in *Turning the tide: How to rescue transatlantic relations* (Paris: EUISS), 55–56.
104 Council of the European Union, "Council conclusions, 8396/21," 13.
105 Simona R. Soare, "Digital divide? Transatlantic defence cooperation on artificial intelligence," *EUISS Policy Brief*, no. 3 (5 March 2020); and Ulrike Franke, "Artificial divide: How Europe and America could clash over AI," *ECFR Policy Brief,* 20 January 2021.

Bibliography

Anghel, Suzana, Beatrix Immenkamp, Elena Lazarou, Jerôme Leon Saulnier, and Alex Benjamin Wilson. "On the path to 'strategic autonomy': The EU in an evolving geopolitical environment." September 2020. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652096/EPRS_STU(2020)652096_EN.pdf.

Bachmann, David, Tobias Bunde, Quirin Maderspacher, Adrian Oroz, Gundbert Scherf, and Kai Wittek. "More European, more connected, more capable:

Building the European armed forces of the future." *European Defense Report*, 2017. https://securityconference.org/publikationen/european-defense-report/.

Barrie, Douglas, Ben Barry, Henry Boyd, Marie-Louise Chagnaud, Nick Childs, Bastian Giegerich, Christian Mölling, and Torben Schütz. "Protecting Europe: meeting the EU's military level of ambition in the context of Brexit." *IISS research paper*, 29 November 2018. https://www.iiss.org/events/2018/11/protecting-europe-brexit.

Bartels, Hans-Peter, Anna Maria Kellner and Uwe Optenhögel (eds). *Strategic Autonomy and the Defence of Europe: On the Road to a European Army?* Dietz, May 2017.

Berger, Roland. "The road to AI – Investment dynamics in the European ecosystem. AI global index 2019." *France Digitale*, 1 January 2019. https://www.rolandberger.com/en/Insights/Publications/The-road-to-AI.html.

Bergmann, Max, James Lamond, and Siena Cicarelli. "The case for EU defense: A new way forward for trans-atlantic security relations." *Center for American Progress,* 1 June 2021. https://www.americanprogress.org/issues/security/reports/2021/06/01/500099/case-eu-defense/.

Borrell Fontelles, Josep. "What's next for European defence?" *European External Action Service*, 7 May 2021. https://eeas.europa.eu/headquarters/headquarters-homepage/98044/what%E2%80%99s-next-european-defence_en.

Borrell Fontelles, Josep. "Why European strategic autonomy matters." 3 December 2020. https://eeas.europa.eu/headquarters/headquarters-homepage/89865/why-european-strategic-autonomy-matters_en.

Borrell Fontelles, Josep. "The Sinatra doctrine: Building a united European front." *Institute Montagne*, 9 September 2020. https://www.institutmontaigne.org/en/blog/sinatra-doctrine-building-united-european-front.

Brands, Hal. "Germany is a flashpoint in the U.S.-China Cold War." *Bloomberg*, 23 February 2021. https://www.bloomberg.com/opinion/articles/2021-02-23/germany-is-a-flashpoint-in-the-cold-war-between-u-s-and-china.

Briani, Valerio, Alessandro Marrone, Christian Moeling, and Tomas Valasek, Tomas. "The development of a European defence technological and industrial base (EDTIB)." *European Parliament, Directorate-General for External Policies of the Union Study*, June 2013. https://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/433838/EXPO-SEDE_ET(2013)433838_EN.pdf.

Brzozowski, Alexandra. "EU leaders caught in either-or argument over European defence strategy." *Euractiv,* 6 October 2021. https://www.euractiv.com/section/defence-and-security/news/eu-leaders-caught-in-either-or-argument-over-european-defence-strategy/.

Burrows, Matthew, and Julian Mueller-Kaler. "Europe's third way." *Atlantic Council*, 14 March 2020. https://www.atlanticcouncil.org/content-series/smart-partnerships/europes-third-way/.

Council of the European Union. "Council conclusions on security and defence." 10 May 2021. https://data.consilium.europa.eu/doc/document/ST-8396-2021-INIT/en/pdf.

Council of the European Union. "Digital sovereignty is central to European strategic autonomy – Speech by President Charles Michel at "Masters of digital 2021" online event." *Press Release,* 3 February 2021. https://www.consilium.europa.eu/en/press/press-releases/2021/02/03/speech-by-president-charles-michel-at-the-digitaleurope-masters-of-digital-online-event/.

Council of the European Union. "Strategic autonomy for Europe – the aim of our generation." *Speech by President Charles Michel to the Bruegel Think Tank,* 28 September 2020. https://www.consilium.europa.eu/en/press/press-releases/2020/09/28/l-autonomie-strategique-europeenne-est-l-objectif-de-notre-generation-discours-du-president-charles-michel-au-groupe-de-reflexion-bruegel/.

Council of the European Union. "Recovery Plan: powering Europe's strategic autonomy." *Speech by President Charles Michel at the Brussels Economic Forum*, 8 September 2020. https://www.consilium.europa.eu/en/press/press-releases/2020/09/08/recovery-plan-powering-europe-s-strategic-autonomy-speech-by-president-charles-michel-at-the-brussels-economic-forum/.

Council of the European Union. "Foreign affairs council conclusions on implementing the EU global strategy in the area of security and defence." 14 November 2016. https://www.consilium.europa.eu/media/22459/eugs-conclusions-st14149en16.pdf.

Council of the European Union. "Implementation plan on security and defence." 14 November 2016. https://www.consilium.europa.eu/media/22460/eugs-implementation-plan-st14392en16.pdf.

Emmott, Robin, and Sabine Siebold. "EU should enable military coalitions to tackle crises, Germany says." *Reuters,* 2 September 2021. https://www.reuters.com/world/europe/eu-must-create-deployable-rapid-reaction-force-borrell-says-2021-09-02/.

European Commission. "European defence fund – Calls 2021." *Factsheet,* 30 June 2021. https://ec.europa.eu/defence-industry-space/edf-calls-2021-factsheet_en.

European Commission. "Coordinated plan on artificial intelligence 2021 review." 21 April 2021. https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review.

European Commission. "Europe's digital decade: Digital targets for 2030." 9 March 2021. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_en.

European Commission. "Action plan on synergies between civil, defence and space industries." *COM(2021) 70 Final*, 22 February 2021. https://ec.europa.eu/info/sites/default/files/action_plan_on_synergies_en.pdf.

European Commission. "Joint communication to the European parliament and the council: The EU's cybersecurity strategy for the digital decade." *JOIN (2020) 18 Final*, 16 December 2020. https://digital-strategy.ec.europa.eu/en/library/eus-cybersecurity-strategy-digital-decade-0.

European Commission. "First six artificial intelligence and blockchain technology funds backed by innovfin raise a total of EUR 700m." 28 October 2020. https://ec.europa.eu/commission/presscorner/detail/en/IP_20_1991.

European Council. "Special meeting of the European council (1 and 2 October 2020): Conclusions, *EUCO 13/20*." 2 October 2020. https://www.consilium.europa.eu/media/45910/021020-euco-final-conclusions.pdf.

European Council. "European council conclusions." *EUCO 217/13*, 19–20 December 2013. https://data.consilium.europa.eu/doc/document/ST-217-2013-INIT/en/pdf.

European Defence Agency. "R&T conference – Impact of disruptive technologies on defence." *Speech by Jiři ŠEDIVÝ, EDA Chief Executive,* 20 April 2021. https://eda.europa.eu/docs/default-source/documents/eda-chief-executive-speech.pdf.

European Defence Agency. "Defence data 2018–2019: Key findings and analysis." 28 January 2021. https://eda.europa.eu/publications-and-data/brochures/defence-data-2018-2019.

European Defence Agency. "2020 Card report: Executive summary." 21 November 2020. https://eda.europa.eu/docs/default-source/reports/card-2020-executive-summary-report.pdf.

European Defence Agency. "Federica Mogherini opens annual conference devoted to unmanned/autonomous systems." 29 November 2018. https://eda.europa.eu/news-and-events/news/2018/11/29/federica-mogherini-opens-annual-conference-focused-on-unmanned-and-autonomous-systems#.

European External Action Service. "Questions and answers: A background for the strategic compass." *Memo,* 3 September 2021. https://eeas.europa.eu/sites/default/files/2021-09-03_-_strategic_compass.pdf.

European External Action Service. "Shared vision, common action: A stronger Europe." *A Global Strategy for the European Union's Foreign and Security Policy*, June 2016. https://eeas.europa.eu/archives/docs/top_stories/pdf/eugs_review_web.pdf.

European Parliament. "AIDA working paper on 'The external policy dimensions of AI' following the AIDA/AFET/SEDE public hearing on 1 and 4 March 2021." *Special Committee on Artificial Intelligence in the Digital Age*, March 2021. https://www.europarl.europa.eu/cmsdata/232488/Working%20Paper%20-%20Joint%20hearing%20on%20the%20external%20Policy%20dimensions%20of%20AI.pdf.

European Parliament. "Artificial intelligence: Questions of interpretation and application of international law in so far as the EU is affected in the areas of civil and military uses and of state authority outside the scope of criminal justice." *2020/2013(INI)*, 20 January 2021. https://www.europarl.europa.eu/doceo/document/TA-9-2021-0009_EN.html.

European Parliament. "European Parliament resolution of 12 September 2018 on autonomous weapon systems." *2018/2752(RSP),* 12 September 2018. https://www.europarl.europa.eu/doceo/document/TA-8-2018-0341_EN.html.

European Union Agency for Cybersecurity (ENISA). "AI cybersecurity challenges: Threat landscape for artificial intelligence." December 2021. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.

Finnish Presidency of the Council. "Food for thought paper by Finland, Estonia, France, Germany, and the Netherlands: Digitalization and artificial intelligence in defence." 17 May 2019. https://eu2019.fi/documents/11707387/12748699/Digitalization+and+AI+in+Defence.pdf/151e10fd-c004-c0ca-d86b-07c35b55b9cc/Digitalization+and+AI+in+Defence.pdf.

Franke, Ulrike. "Artificial intelligence diplomacy: Artificial intelligence governance as a new European Union external policy tool." *European Parliament*, June 2021. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662926/IPOL_STU(2021)662926_EN.pdf.

Franke, Ulrike, and Jose Ignacio Torreblanca. "Geo-tech politics: Why technology shapes European power." *ECFR*, 15 July 2021. https://ecfr.eu/publication/geo-tech-politics-why-technology-shapes-european-power/.

Franke, Ulrike. "Artificial divide: How Europe and America could clash over AI." *ECFR Policy Brief,* 20 January 2021. https://ecfr.eu/publication/artificial-divide-how-europe-and-america-could-clash-over-ai/.

German Presidency of the European Council. "Independent, inclusive and innovative: Four goals of the German Presidency for the digital sector." *EPC*, December 2020. https://www.eu2020.de/eu2020-en/news/article/digitalziele-eu2020/2405548.

Grevi, Giovanni. "Fostering Europe's strategic autonomy – A question of purpose and action." *EPC*, December 2020. https://epc.eu/content/PDF/2020/Final_Paper_Purpose_and_Action_Layout_JF_II__1_.pdf.

Herszenhorn, David M. "Biden's top security adviser sees strong transatlantic alliance (and no jumping in lakes)." *Politico*, 8 October 2021. https://www.politico.eu/article/jake-sullivan-biden-national-security-transatlantic/.

Herszenhorn, David M. "German, Dutch diplomats urge stronger NATO-EU ties." *Politico,* 20 May 2021. https://www.politico.eu/article/german-dutch-paper-urges-stronger-nato-eu-ties/.

Hobbs, Carla (ed). "Europe's digital sovereignty: From rulemaker to superpower in the age of US-China rivalry." *ECFR,* 30 July 2020. https://ecfr.eu/publication/europe_digital_sovereignty_rulemaker_superpower_age_us_china_rivalry/.

IISS. "Cyber capabilities and national power: A net assessment." 28 June 2021. https://www.iiss.org/blogs/research-paper/2021/06/cyber-capabilities-national-power/.

Ilves, L., and A.-M. Osula. "The technological sovereignty dilemma – and how new technology can offer a way out." *European Cybersecurity Journal* 6, no. 1 (2020): 24–35.

Izsak, Kincsö, Maialen Perez, Henning Kroll, and Sven Wydra. "Advanced technologies for Industry – EU Report Technological trends and policies." *Joint Research Center,* November 2020. https://ati.ec.europa.eu/reports/eu-reports/eu-report-technological-trends-and-policies.

Ko, Ryan. "Cyber autonomy: Automating the hacker – self-healing, self-adaptative, automatic cyber defence systems and their impact on society, industry, and national security." In *Emerging Technologies and International Security: Machines, the State, and War,* edited by Reuben Steff, Joe Burton, and Simona R. Soare, 173–191, London: Routledge, 2020.

Kortunov, Andrey. "Russian perspective on the challenges to the European project." *Russian International Affairs Council*, 9 July 2021. https://russiancouncil.ru/en/analytics-and-comments/analytics/russian-perspective-on-the-challenges-to-the-european-project/.

Kott, Alexander et al. "Autonomous intelligent cyber-defense agent (AICA) reference architecture release 2.0." *US Army Research Laboratory*, September 2019. https://arxiv.org/pdf/1803.10664.pdf.

Le Gleut, Ronan, and Hélène Conway-Mouret. "European defence: The challenge of strategic autonomy." *French Senate Report No. 626 Prepared for the Extraordinary Session of 2018–2019,* 3 July 2019. http://www.senat.fr/rap/r18-626-2/r18-626-20.html.

Lippert, Barbara et al. (eds.) "European strategic autonomy: actors, issues, conflicts of interests." *SWP Research paper,* March 2019. https://www.swp-berlin.org/publications/products/research_papers/2019RP04_lpt_orz_prt_web.pdf.

Martineau, Kim. "Deep-learning models code more like humans." *Control Engineering,* 23 April 2021. https://www.controleng.com/articles/deep-learning-models-code-more-like-humans/.

Ministère des Armées. "Strategic update 2021." February 2021. https://www.defense.gouv.fr/content/download/605304/10175711/file/strategic-update%202021.pdf.

Molenaar, Arnout. "Unlocking European defence. In search of the long overdue paradigm shift." *IAI,* 22 January 2021. https://www.iai.it/en/pubblicazioni/unlocking-european-defence.

Munich Security Conference. "Munich security index 2021: Appendix to the Munich security report 2021." *Additional Survey Results and Analysis*, February 2021. https://securityconference.org/en/publications/munich-security-index-2021/.

Mustasilta, Katariina. "Preventing our way back to friendship? Conflict prevention and the future of transatlantic relations." In *Turning the tide: how to rescue transatlantic relations*, edited by Simona R. Soare, Paris: EUISS, 2020.

Official Journal of the European Union. "Regulation (EU) 2021/697 of the European parliament and of the council of 29 April 2021." *European Defence Fund and Repealing Regulation (EU),* 29 April 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32021R0697&from=EN.

Pavel, Claudiu, and J. Huisman. "Critical raw materials for strategic technologies and sectors in the EU: A foresight study." *Joint Research Center,* 14 September 2020. https://ec.europa.eu/docsroom/documents/42881.

Pavel, Claudiu, and Evangelos Tzimas. "Raw materials in the European defence industry." *Joint Research Center,* 2016. https://publications.jrc.ec.europa.eu/repository/handle/JRC98333.

Pfeiffer, Simon, and Randolf Carr. "Trust not found: A European survey on digital (dis)trust." *Munich Security Conference brief*, March 2021. https://securityconference.org/assets/02_Dokumente/01_Publikationen/MunichSecurityBrief_Error404_TrustNotFound.pdf.

Sarker, Iqbal H., Hasan Furhad, and Raza Nowrozy. "AI-driven cybersecurity: An overview, security intelligence modelling and research directions." *SN Computer Science* 2 (2021). https://doi.org/10.1007/s42979-021-00557-0.

Soare, Simona R. "Digital divide? Transatlantic defence cooperation on artificial intelligence." *EUISS Policy Brief*, 5 March 2020. https://www.iss.europa.eu/content/digital-divide-transatlantic-defence-cooperation-ai.

Soare, Simona R. "European defence and AI: Game-changer or gradual change?" *RSIS Commentary,* 24 March 2021. https://www.rsis.edu.sg/wp-content/uploads/2021/03/CO21051.pdf.

Soare, Simona R. "European military AI: Why regional approaches are lagging behind." In *Global Strategic Perspectives on Military AI*, edited by Michael Raska and Zoe Stanley-Lockman, Singapore: Routledge, 2022.

Soare, Simona R. "Partners in need or partners in deed? How EU-NATO cooperation shapes transatlantic relations." In *Turning the Tide: How to Rescue Transatlantic Relations*, edited by Simona R. Soare, Paris: EUISS, 2020. https://www.iss.europa.eu/sites/default/files/EUISSFiles/Transatlantic%20relations%20book.pdf.

Soare, Simona R. "Politics in the machine: The political context of emerging technologies, national security, and great power competition." In *Emerging Technologies and International Security: Machines, the State, and War*, edited by Reuben Steff, Joe Burton, and Simona R. Soare, London: Routledge, 2020.

Soare, Simona R., and Fabrice Pothier. "Leading edge: Key drivers of defence innovation and the future of operational advantage." *IISS Research Paper,* November 2021. https://www.iiss.org/blogs/research-paper/2021/11/key-drivers-of-defence--innovation-and-the-future--of-operational-advantage.

Tanguy, Jean-Marc. "France launches first Syracuse IV telecommunications satellite." *Jane's*, 26 October 2021. https://www.janes.com/defence-news/news-detail/france-launches-first-syracuse-iv-telecommunications-satellite.

Techau, Jan. "Why the EU can't do security and defence." *EUObserver*, 23 October 2019. https://euobserver.com/opinion/146369.

Thomson, Robert, Christian Lebiere, and Drew Cranford. "Achieving active cybersecurity through agent-based cognitive models for detection and defense." *United States Military Academy of West Point*, 2021. https://www.aica2021.org/

wp-content/uploads/2021/03/Thomson-AICA-2021-Achieving-Active-Cybersecurity-through-Agent-Based-Cognitive-Models-for-Detection-and-Defense.pdf.

Tigner, Brooke. "EU's proposed rapid reaction entry force faces many hurdles." *Jane's*, 11 May 2021. https://www.janes.com/defence-news/news-detail/eus-proposed-rapid-reaction-entry-force-faces-many-hurdles.

Von der Leyen, Ursula. "Shaping Europe's digital future: op-ed by Ursula von der Leyen, President of the European Commission." *European Commission*, 19 February 2021. https://ec.europa.eu/commission/presscorner/detail/en/AC_20_260.

Voo, Julia, Irfan Hemani, Simon Jones, Winnona DeSombre, Daniel Cassidy, and Anina Schwarzenbach. "National cyber power index 2020: Methodology and analytical considerations." *Harvard Belfer Center for Science and International Affairs Report*, September 2020. https://www.belfercenter.org/sites/default/files/2020-09/NCPI_2020.pdf.

Youngs, Richard. "The EU's strategic autonomy trap." *Carnegie Europe*, 8 March 2021. https://carnegieeurope.eu/2021/03/08/eu-s-strategic-autonomy-trap-pub-83955.

Zandee, Dick, Bob Deen, Kimberley Kruijver, and Adája Stoetman. "European strategic autonomy in security and defence: Now the going gets tough, it's time to get going." *Clingendael Report*, December 2020. https://www.clingendael.org/sites/default/files/2020-12/Report_European_Strategic_Autonomy_December_2020.pdf.

Zuiderwijk, Anneke, Yu-Che Chen, and Fadi Salem. "Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda." *Government Information Quarterly* 38, no. 3 (July 2021): 1–19. https://doi.org/10.1016/j.giq.2021.101577.

Whyte, Christopher. "Poison, persistence, and cascade effects: AI and cyber conflict." *Strategic Studies Quarterly* 14, no. 4 (2020): 18–46.

Whyte, Christopher. "Scenario 2—AI and insecurity for all: The future of cyber conflict." *Alternate Cybersecurity Futures*, September 2019. https://www.atlanticcouncil.org/wp-content/uploads/2019/08/Alternate-Cybersecurity-Futures-FINAL.pdf.

Wu, Yirui, Dabao Wei, and Jun Feng. "Network attacks detection methods based on deep learning techniques: A survey." *Hindawi Security and Communication Networks* (2020). https://doi.org/10.1155/2020/8872923.

5 The middleware dilemma of middle powers

AI-enabled services as sites of cyber conflict in Brazil, India, and Singapore

Arun Mohan Sukumar

Introduction

Although no accepted definition of the term exists, "middle powers" includes those countries that exercise a high degree of economic sovereignty,[1] and strongly influence regional developments through their political, economic, cultural, or military capabilities. Middle powers may also influence rule-making in specific domains of global governance. Notably, these countries tend to favor the *status quo* of the liberal international order, preferring to engage multilateral regimes and seek accommodations from them as needed, rather than challenging those regimes or the broader, hegemonic interests that underpin them.[2] "Middle powers" encompasses both developed and developing countries. Indeed, the term is as much a reflection of states' material capabilities as it is of "normative and behavioral criteria."[3] Middle powers aspire to maintain and elevate their international status. Technological leapfrogging – the adoption by developing countries of frontier technologies for governance, skipping in the process older-generation technologies that are either resource-intensive or unscalable – figures prominently in this pursuit of status.[4] Leapfrogging, whether through the adoption of GSM telephony, IPv6, or 5G, has not just been viewed as a sustainable path to economic prosperity. It is also a totem of empowerment for middle powers, allowing them to participate on equal footing with Great Powers on Research & Development, standard-setting, and application of new technologies for their own requirements.

Artificial Intelligence and Machine Learning (AI/ML) represent one such frontier technology. Since the publication of the world's first national AI strategy in 2017 by Canada – a self-described "middle power"[5] – no less than 54 countries have either announced or are in various stages of declaring their own national strategies.[6] Several of these strategies, especially those drawn by middle powers, emphasize the importance of AI in leapfrogging hurdles to the delivery of services in sectors such as healthcare, education, transportation, and e-commerce. In particular, they underline the need to take advantage of vast troves of data generated by their population to train

DOI: 10.4324/9781003284093-7

predictive algorithms and develop ML models for the efficient delivery of said services. Given the costs and limited availability of high-skilled workers to train healthcare professionals, lawyers, and educators, among others, emerging markets may seek to replace them with AI-enabled 'bots.'

A review of national AI strategies suggests states may pursue one of two paths to promote innovation and adoption of AI-enabled services. Some countries may choose to "make public datasets available" for market players to develop proprietary AI-enabled services, and concurrently, set up regulatory sandboxes to pilot those products.[7] Other states may opt to leverage their Digital Public Infrastructure (DPI), i.e., public technical standards and protocols that allow third parties to access personal data of citizens or anonymized data, with clearly defined guidelines on the nature of data that can be shared, the duration of data sharing, as well as permissible use-cases.[8] Such infrastructure, which can be considered "middleware," then becomes the conduit for access of data to train AI/ML models. The "middleware" model is likely to be preferred by many states as it allows for more granular, revocable, and regulated data-sharing, rather than one-time access to public databases. Middle powers in particular may condition the use by market players of such middleware on their affordable provision of AI-enabled services in sectors such as health and education.

Examples of such middleware infrastructure developed by middle powers include protocols that standardize data-sharing (for example, Personal Health Records in Japan),[9] common and interoperable railroads for digital transactions (PayNow in Singapore,[10] Unified Payments Interface in India,[11] etc.), or unique digital identifiers, often validated by biometric markers (Aadhaar in India,[12] MOSIP in Philippines,[13] e-ID system in Estonia,[14] etc.). The leveraging of middleware DPIs towards AI-enabled services is in the early stages of conceptualization and implementation at the time of writing. Nevertheless, it is highly likely that middle powers that have made considerable advancements in building such infrastructure – Australia, India, Brazil, Norway, Japan, South Africa, and Singapore, to name a few – will rely on them to promote innovation and at-scale adoption of AI-driven services. Indeed, the high degree of standardization in data collection, labelling, and sharing through middleware DPIs makes it easier to develop and train ML models. DPIs also allow regulators to calibrate the nature of data collected, allowing in some cases for the sharing of scarce personal and population-level information, and minimizing the collection of potentially harmful information in others. These reasons make it probable that public, middleware infrastructure will be used to develop AI-based applications and services across the world. For instance, it is worth highlighting how the COVID-19 pandemic has accelerated the adoption of AI-enabled diagnostics that rely on data from contact-tracing applications or chest imaging databases built by states.[15] This trend will likely continue in the future, especially in middle powers that have created centralized national databases faster than advanced economies.

The development of AI-enabled products and services that rely on such middleware infrastructure also creates a dilemma for states. DPIs are heavily reliant on Application Programming Interfaces (APIs) that lay down technical specifications for data collection and sharing. In some cases, the middleware in question is nothing but an API, i.e., a few lines of code that allow public and private players to connect with each other and share data. Payment railroads are a common example of such APIs. In other instances, APIs are necessary to allow external applications to connect to the public infrastructure and retrieve data from its servers. With respect to AI-enabled services or products, APIs will be key to ensuring that ML models developed by third parties have access to the right data parameters in order to train their algorithms. Essential as they are to the adoption and scaling of digital infrastructure, APIs are also vulnerable to sophisticated cyber attacks. By their very nature, APIs are designed to facilitate the seamless integration of third-party applications with databases, cloud services, virtual networking functions, etc. The emphasis on ease of access has often come at the cost of secure API design.[16] In the case of AI-enabled services that run on public datasets or middleware infrastructure, this vulnerability is compounded by the fact that there is minimal human supervision of the interaction between the service and the user. As a result, attacks on AI-enabled services through APIs could seriously impair both the availability of the service as well as its predictive accuracy and effectiveness. Attacks against the availability or integrity of AI-enabled services may thus have the effect of undermining public confidence and trust in them, especially in developing countries and middle powers.

To be sure, APIs are potent vectors for cyber attacks wherever they are deployed. With API 'calls' comprising over 80% of the Internet's platform-led traffic, such vectors are omnipresent.[17] The role of APIs in linking DPI to AI-enabled services, however, makes them an even more lucrative target for state and non-state actors who seek not only entry into critical national databases, but also disrupt autonomous and minimally supervised platforms that provide essential services.

This chapter highlights how AI-enabled services running on public, digital infrastructure could emerge as vectors of cyber conflict. States face a difficult choice between opening up national databases or other public infrastructure to third parties in order to promote AI innovation atop its data and risking not only the security of those databases but also that of critical, even lifesaving, services. The "middleware dilemma" is most acute for middle powers, especially large emerging markets that are undergoing rapid industrialization but find themselves constrained by resources or capital to provide essential services at scale. As a result, they are compelled to turn towards digital or digital-enabled services, including AI services.

The chapter is organized as follows: section two explores the 'middleware dilemma' in detail, highlighting security risks associated with API deployment, and their increasing role in the training of AI/ML modes. Section three outlines efforts by three middle powers – Brazil, Singapore, and

India – to make available public data and digital infrastructure through APIs to promote AI innovation in their healthcare sector. Drawing on these national models and strategies, Section four presents an overview of threats and vulnerabilities faced by states in securing such services and their underlying infrastructure.

APIs and the "middleware dilemma"

The internet is witnessing unprecedented 'API-fication.' APIs are lines of code that allow software to communicate with each other, and thus facilitate greater connectivity and interoperability among the network, data, and application layers that make up cyberspace. With digital environments becoming increasingly heterogenous – most businesses and enterprises today delegate routing, data processing, and even cybersecurity functions to virtual networks and cloud services located halfway across the world – APIs have become crucial to the smooth functioning of critical internet services. Through their enabling role in the retrieval and processing of data at the application/device level, APIs have also been instrumental in realizing the "platform economy," as it is known today.[18] For the same reason, APIs have been key to the rapid expansion and proliferation of federated and centralized databases across the world. The DPI developed by middle powers such as Australia, Brazil, and India, to name a few countries, too depend on APIs for their implementation and use. Indeed, regulatory tools such as the 2018 Revised Payments Directive (PSD2) in the European Union and the 2020 Consumer Data Right Act in Australia require even private actors to provide standardized APIs so that user data is interoperable and seamlessly accessible by all authorized parties.

Despite their popularity, however, API security still leaves much to be desired. APIs have earned notoriety in recent years as attack surfaces for data breaches, identity theft and account takeovers, ransomware injections, IoT exploitation and DDoS attacks, among others.[19] One estimate suggests API attacks will emerge as the leading vector of cyber attacks by 2022. The security considerations involving APIs are three-fold: first, the widespread use of APIs results in a crowding of digital networks and infrastructure by third parties, making it difficult to manage or monitor the proliferating endpoints. Indeed, network administrators have no 'over-the-horizon' visibility with respect to third-party applications or devices that are constantly pinging their databases or infrastructure with API calls. Second, APIs may be developed by actors across the network, but there are no clear frameworks for accountability and remedial action in case of API-enabled attacks.[20] The poor maintenance and updating of APIs has even contributed to the phenomenon of 'zombie' APIs that continue to be functional (and potentially leak data) although their developers have long abandoned their active use.[21] Finally, a culture of data maximalism – "when in doubt, collect" – pervades API

design, with the result that API attacks often result in "excessive data exposure"[22] of users. A design culture favoring ease of access has also resulted in the neglect of security evaluations in the development cycle, although there is more awareness among API programmers today than even the recent past.

On account of these factors, threats to API security have risen in severity and sophistication. The Open Web Application Security Project Foundation's (OWASP) annual 'Top 10' rankings of API security threats – considered a benchmark among market players and cybersecurity researchers alike – has consistently identified the following as high-priority concerns:[23]

a Code injection, i.e., pinging the API with malicious code that allow unauthorized actors to retrieve, manipulate, or destroy user data;
b Broken authentication, i.e., the use of APIs for credential stuffing or brute force attacks that permit malicious actors from taking control of user accounts associated with a service or application;
c Man-in-the-Middle attacks that take advantage of poor encryption protocols (at rest or in transit) to retrieve highly sensitive user details;
d Insecure design of APIs that lean on legacy methods to recover user credentials, retrieve data, generate error messages, etc., without adequate threat modeling.

Given the nature and gravity of such threats, the use in particular of APIs that allow third parties to connect and retrieve information from middleware infrastructure presents a major cybersecurity concern for states. In the case of many middle powers, especially large developing countries, the government plays an important role in shaping the digital economy. The state in question may want to share data with market players in a bid to boost private innovation. While APIs present a relatively easy and seamless way for many states to create middleware that collects data or retrieves it from existing databases, they must balance such convenience against the risk of losing highly sensitive information. In some instances, only states have the legal imprimatur to collect certain types or categories of data from citizens, and the possibility of leakage or unauthorized exposure of such data (for example, biometric or health data) to third parties through APIs is high.

The concern that APIs may be vectors for cyber attacks and indeed, cyber conflict, is compounded in the case of AI-enabled services. States may allow the use of APIs specifically to promote AI/ML innovation on public databases or infrastructure in a number of ways. Governments could lay down specifications and protocols for retrieving data either from databases or directly from citizens. Once data is collected through such traditional, "dumb" APIs, it is left to the third party to anonymize the data and use it to train their proprietary ML models. Alternatively, states could provide "clean rooms" or closed environments where personal and population-level data is anonymized and training models built without actual transfer of data. Such a model relies

on advancements in secure multi-party computation that allows for training of algorithms without having to share private data.[24] And finally, states could open up their infrastructure to third party Machine Learning APIs (ML APIs) that are used by start-ups, enterprises, and public agencies alike. Indeed, this third option may emerge as a popular one for many market players who do not themselves have the capacity or resources to train ML models, but have innovative AI-enabled services to offer. The widespread adoption of cloud computing has boosted the popularity of ML-as-a-Service: AI-enabled products and services have increasingly begun to offload data processing and training of algorithms to cloud-based ML services such as Google, AWS, and Azure. Their APIs perform a number of critical functions, providing both 'off-the-shelf' and customizable neural networks to third-party applications. For start-ups that want to train their own algorithms on cloud-based services, ML APIs are invaluable for data labeling, maintaining registries of training models, and for periodic audit of those models.[25]

These methods of using APIs to facilitate third party access and innovation in AI-enabled services are not without risks. The security considerations and threats involving APIs in general have already been documented in this section, and need not be repeated here. Such threats are, however, more pronounced with respect to the use of APIs to train ML models. With greater volumes of data being called by APIs for training purposes, they become lucrative targets for state and non-state adversaries. A major concern with the use of traditional and ML APIs is their handling of data, and the measures taken by states as well as private actors to not only anonymize training data but also minimize risks of subsequent de-anonymization.[26] In many developing countries, judicial and regulatory capacity to address de-anonymization risks breaches may be limited, as a result of which its resolution could be entirely dependent on voluntary, technical steps taken by market players. Without effective safeguards to prevent de-anonymization, APIs could be exploited by adversaries to capture highly sensitive details from training data about the population.

ML APIs have surged in popularity, especially in the aftermath of the COVID-19 pandemic. With businesses and NGOs moving their operations online, the demand for AI-enabled audio/video, text-based, and Natural Language Processing (NLP) services has increased significantly. The health sector has seen perhaps the biggest transformation during this period, as witnessed by the move towards predictive diagnostics and 'health bots' that perform remote consultation. The rapid rise and adoption of ML APIs raise the concern that their software design may sidestep security considerations in favor of scale and ease of access. It is not simply the secure design of ML APIs that matter, but also their use by developers or services who are new to using 'off-the-shelf' tools for training ML models. As Wan et al. note, ML API misuses have already become commonplace, because start-ups or businesses are not fully aware of attributes of ML tools offered by cloud services like

Amazon or Google.[27] Their study of 360 applications that relied on ML APIs found that developers routinely called the wrong API – for e.g., many apps confused "image classification" APIs with "object detection" APIs, the latter being used to identify objects within an image – which affected the accuracy and effectiveness of the service.[28] Additionally, many developers also interpreted the predictive results delivered by ML APIs incorrectly, mistaking probabilistic assessments for binary ('yes' or 'no') results.[29] A poorly understood and utilized ML API ecosystem is ripe for exploitation and disruptive cyber attacks.

As more applications and services rely on ML APIs to retrieve and train data through DPI, states will thus be confronted by the challenge of securing that data against API design flaws, improper use, and exploitation. Training models not only require large datasets, but are also in need of constant updates both to the ML APIs as well as the data itself. As Chen et al. note, the predictive performance of ML APIs can grow "significantly worse over time" even when they rely on the same datasets.[30] Routine updates both to the API (by the cloud-based provider) and the training data are crucial to the model's effective performance. Unfortunately, such a highly dynamic environment also increases the chances for MITM attacks that may be carefully disguised as API updates or requests for new data. Cloud service providers have been criticized in the past for considering security as an "externality," and shifting the loss from cyber attacks on to the users.[31] If they adopt the same approach with respect to ML APIs, especially those that 'call' public infrastructure, states may be constrained to address cyber attacks on their infrastructure and AI-enabled services quickly and effectively. A 2017 review of iOS and Android developer guidelines found many aspects of application-layer security on these platforms to be insufficient or only partly aligned to OWASP standards.[32] With no human supervision of interactions between AI-enabled services and their users, similar vulnerabilities in ML APIs could be exploited to disastrous consequences.

In summary, vulnerabilities associated with traditional and ML APIs could result in the misuse, manipulation, and even denial of AI-enabled essential services that rely on them. Given deficiencies in secure API design, and in many instances, their poorly understood application with respect to core functions, API-driven middleware could be prime targets of strategic adversaries in the event of conflict. Given these concerns, it is worth examining the different approaches of middle powers with respect to the adoption of APIs for AI-enabled services in critical sectors, and the possible security repercussions of those API-led models.

Case studies: Brazil, India, and Singapore

The critical sector of healthcare has been identified by several states, including middle powers, as ripe for technology leapfrogging. Emerging markets

and developing countries with large populations have historically struggled to train medical professionals whether in the field of diagnostics or healthcare services. With a view to address the lack of skilled professionals, governments have turned to digital healthcare services. Furthermore, the COVID-19 pandemic has catalyzed rapid advancements in digital health, including in the development of AI-enabled diagnostics and services. However, digitizing sensitive health data of populations and rendering them accessible to third parties – via API-based middleware – raises the possibility of such information being compromised or corrupted by adversaries.

The following section outlines recent and ongoing efforts by three middle powers – Brazil, Singapore, and India – to make available public health data via APIs for external developers, including of AI-enabled applications in the sector. In particular, it emphasizes those historical and institutional reasons why these states have chosen to pursue three different approaches to using APIs for facilitating third-party access to healthcare databases.

Brazil

Background

Among developing countries and middle powers, Brazil stands out as a pioneer in the 'informatization' of national healthcare services. Brazil's Unified Health System (SUS), a universal healthcare program established in 1988, is among the largest of its kind in the world.[33] Since 1993, Brazil has created several specialized national databases pertaining to vaccinations, cancer screening and treatments, infectious disease surveillance, movement of restricted drugs, patient visits, and social security benefits. However, it has struggled to digitize these databases and make them interoperable across sectors and healthcare providers.[34] Consequently, private players have stepped in to build their own algorithms for data retrieval and linkage from these databases. Given some of these databases have no anonymization features,[35] the involvement of private actors has raised concerns around privacy and cybersecurity. Although an 'e-SUS' platform has been in existence since 2014 to collect primary healthcare data and population-level indicators, the development of this platform has been hampered by a lack of training in data-entry among healthcare workers as well as "bureaucratization of their work process."[36] Another major challenge in digitizing and consolidating such data has been the lack of a unique identity program in Brazil.[37] And finally, the absence until 2020 of an overarching data protection legislation meant there were no general legal or policy measures governing the handling of sensitive health data. As a result of all these factors, Brazil's expansive policy infrastructure on healthcare and social security has historically been challenged by a skeletal digital infrastructure with no "semantic and technological standardization" for data.[38] However, this scenario has changed dramatically in the aftermath of the COVID-19

pandemic, whose precipitation of the demand for digital healthcare services appears to have been seized by both government and private actors.

The role of APIs in the digitization of healthcare

With the onset of the COVID-19 pandemic, Brazil re-oriented the implementation of three key policy instruments – the National Digital Health Strategy, 2020–2028 (NDHS), National Health Data Network (RNDS) (2020), and the National Artificial Intelligence Strategy, 2021 (NAIS) – to mitigate the spread of the coronavirus and manage the treatment of those infected. These policies were in advanced stages of consultations well before the pandemic, but Brazilian regulators were compelled by the coronavirus' rapid spread to digitally unify various elements of the health system in a bid to address COVID-19 surveillance, immunization, adequate availability of hospital facilities, testing records, etc. For example, the RNDS was initially supposed to be rolled out as a pilot project in a single Brazilian province in March 2020, but was repurposed to "receive and share information [across the country] that could help [the government] control" the pandemic.[39] Similarly, the national health strategy emphasized the interoperability of data across healthcare providers to help tackle together the spread of COVID-19.[40] Finally, the national AI strategy declared that health would be one of the first sectors to see the roll out of AI-driven pilot and implementation projects.[41]

The RNDS in particular is slated to play a critical role in the standardization and interoperability of health data in Brazil. The NDHS declares the eventual objective of the RNDS to be the creation of an ecosystem where the "SUS, public and private healthcare organizations, technology companies, research centers, universities and other stakeholders share data […] well as exercise, test and evaluate new models, patterns, technologies and design."[42] In July 2020, Brazil made the submission of SARS-CoV-2 diagnostic tests to the RNDS – whether conducted by public or private laboratories – mandatory.[43] Following this legal measure, the SUS created "accrediting systems and technical documentation" in its 'DATASUS' platform to facilitate such submission and data sharing.[44] The technical documentation in question referred to a set of API standards. At the time of writing, the Brazilian Ministry of Health has expanded the suite of APIs available in DATASUS, and includes those that not only allow for the sharing of test data, but also the sharing of clinical studies results, immunization data, pharmacy inventories, and primary healthcare data into RNDS.[45] The ministry's Coronavirus-SUS app, used for contact tracing, relied on the Google/Apple Exposure Notification (GAEN) API developed jointly by the two companies.[46] Although it remains possible to export data from government websites or health applications directly, the Brazilian government has strongly encouraged the use of these APIs[47] over other channels, creating the basis for a digital health architecture that is heavily reliant on middleware.

AI-enabled services and future plans

The Brazilian ordinance of August 2020 that established the RNDS offers an insight into the role of APIs in promoting AI-driven innovation in the country's health sector. The ordinance attempts to promote "interoperability" in:[48]

a Information models, i.e., "conceptual and contextual human representation" of data;
b Computational models, i.e., data structures as programmed in a computing language; and
c "Semantic" and "syntactic" data models, i.e., human and computational representations respectively of "classifications, taxonomies, and ontologies" and other information models relevant to the sector.

From Brazil's detailed and carefully crafted attempts to introduce standardization and interoperability in electronic health records, it is amply clear the country's regulators do not see APIs simply as a quick fix towards digitizing the sector. Instead, Brazil's recent national strategies on digital health and AI, as well as a slate of pandemic-era policies, appear to signal the creation of an API-centric middleware ecosystem that facilitates the sharing of personal and non-personal data, and in turn, promotes innovation in AI-enabled services. Brazil's National Health Information and Informatics Policy (PNIIS), introduced in July 2021, specifically call for the use of AI to meet the needs of healthcare professionals and researchers.[49] The 'Conecte-SUS' app, which provides users with a longitudinal record of their clinical history that can be shared with healthcare providers and researchers via DATASUS APIs, has already been earmarked by local governments as a DPI to promote AI-enabled innovation.[50] Several multistakeholder pilot projects on predictive COVID-19 diagnostics have already been implemented in Brazil, although it is unclear at the time of writing whether they have relied on APIs or single-site data. In any event, the 'API-fication' of Brazil's digital health sector appears to be a deliberate and ambitious strategy to ensure public and private agencies can rely on large volumes of data to train and develop ML models in primary healthcare and diagnostics.

Singapore

Background

As a middle power with outsize ambitions to shape normative and material outcomes in cyberspace, Singapore has long sought the comprehensive digitization of key sectors of domestic governance. GovTech, a specialized agency established in 2016 to catalyze the digital transformation of Singapore's public sector and user-facing services, makes a credible claim to be "the first of its

kind" in the world.[51] Among its other responsibilities, GovTech is responsible for the country's "Strategic National Projects" which includes user- and business-facing platforms to access government services, the national digital identity program (Singpass), the country's unified payment gateway (PayNow), and CODEX, a technology stack to standardize the development of applications and handling of data across the country's private and public sectors.[52] While Singapore thus pioneered the development of several DPIs, it has moved cautiously with respect to the digitization of its healthcare sector. The fact that Singapore's "worst cyber attack" implicated its national SingHealth system, may have been a contributing factor.[53] The country's digital health policies initially monitored standalone products and services for quality assurance, risk attributes, and adverse event reporting, and it was only in 2020 that the government sought to address questions regarding the integrity and security of health data.[54] As in the case of Brazil, the COVID-19 pandemic catalyzed the creation of legal and technical frameworks on health data in Singapore. The Regulatory Guidelines on Software Medical Devices, issued in April 2020, underline application-layer security concerns similar to those identified by OWASP. The Guidelines call on software developers to ensure, among others:[55]

a Use of proper authentication protocols, both at the device and API levels;
b Development of "layered authorization models" to differentiate privilege levels for users and devices;
c Encryption for data at rest and transit; and
d Deployment of network monitoring and intrusion detection systems.

Notably, the Guidelines also specify regulatory requirements for AI-enabled medical devices and services. AI/ML services that rely on 'static' datasets as well as continuous learning are required to submit descriptions of data attributes, labels, training models, built-in audit processes, and security features, prior to their registration with Singapore's Health Sciences Authority.

The role of APIs in the digitization of healthcare

Singapore has conceptualized its DPI as platforms, and not specific products, believing the latter to be an impediment to at-scale delivery of services.[56] Consequently, APIs have played a prominent role in connecting these middleware platforms to market actors and end-users. Singapore's "platform-as-a-service" model is made possible by the presence of an "engagement layer" of APIs that connect various agencies and institutions within government.[57] Indeed, Singapore's (now) Chief Digital Officer has characterized some of these APIs as "whole-government APIs."[58] They allow, for instance, businesses to obtain licensing or regulatory approvals from multiple bureaucracies through a single application. Similarly, through the integration of the Singpass system across various government platforms, the Singaporean citizen can avail of any public service through her digital ID.

Through the creation in 2017 of a centralized API Exchange (APEX), Singapore brought its 'whole-government' APIs under an umbrella framework. APEX allows government agencies to share data with each other as well as the broader public, allowing, for instance, private services to retrieve user data previously authenticated by the state, or citizens to submit governance proposals that are then channeled to the appropriate entity.[59] Given its extensive interface with government portals and sensitive data, agencies and third parties are required to undergo a training session and test application-layer security protocols before APEX onboarding is complete.[60]

More pertinent to the context at hand, Singapore has also created a portal called 'data.gov.sg' that offers third parties access to public datasets through APIs.[61] Set up in 2011, the portal was criticized in its initial years for being a "data dump" of files in PDF and CSV format that had to be manually downloaded.[62] In recent years, it has undergone a comprehensive transformation, and while data files may still be downloaded, it is through APIs that the outside world engages with 'data.gov.sg.' Most importantly, the portal also makes available APIs that allow for the retrieval of real-time data in such domains as meteorology and transport.

'Data.gov.sg' hosts over 100 datasets pertaining to health.[63] These include data on infectious disease prevalence, incidence of cancer among the population, preventive health screening results, prevalence of so-called 'lifestyle' diseases such as hypertension, diabetes, cholesterol and obesity, hospital facilities and physicians by secondary and tertiary sectors, immunization statistics, and of course, COVID-19-related information. It is worth noting here that many health-related datasets have been made available through APIs following the onset of the coronavirus pandemic, although they have been in existence for years. The rapid onboarding of health data for third party access, combined with Singapore's overall vision and concerted push to promote "open data" governance through APIs suggests the government is heavily leaning on middleware-driven innovation in healthcare services.

AI-enabled services and future plans

In October 2021, Singapore published AI in Healthcare Guidelines (AIHGle) that offer non-binding recommendations to developers and adopters for the "safe implementation" of AI-enabled medical devices.[64] While the Guidelines devote their attention mainly to questions of fairness and explainability of algorithmic decision-making, as well as end-user communication about the working of such AI-enabled devices, security considerations also figure prominently in the document.

The AIHGle identifies both "data risks" and "algorithmic risks" with respect to the security of the AI-enabled service.[65] To mitigate data risks, the Guidelines recommend safeguards against unauthorized access (through APIs or otherwise) to testing, training, and clinical data. The AIHGle also suggests de-identifying personal data where possible, and where "individual

characteristics need to be retained," using techniques such as "data masking, pseudonymization, or data perturbation."[66] To prevent re-identification, developers are encouraged to keep access logs and apply, where possible, techniques such as secure, multi-party computation. The document acknowledges algorithmic risks, i.e., security concerns pertaining to learning and implementation lifecycles, are more accentuated in AI/ML services that rely on "continuous learning" through dynamic and real-time data flows. In such cases, implementers are encouraged to monitor abnormal algorithmic behavior caused by "maliciously introduced data" or manipulations at the end-user level.[67] The AIHGle places much emphasis on human intervention in the implementation process. Implementers of AI-enabled services should have "self-validation" or fail-safe mechanisms that trigger human intervention when baseline performance of the algorithm is affected and even "contingency plans" that "include shutting down the AI device and switching to analogue protocols."[68]

Although the Guidelines are notable for their level of detail and specifications with respect to cybersecurity as well as algorithmic decision-making, it is unclear how its non-binding recommendations will be enforced by the Singaporean government. The AIHGle recommends developers and implementers enter into Service Level Agreements (SLAs) that demarcate their respective responsibilities for the training and implementation lifecycle.[69] Given the Guidelines are only a few months old at the time of writing, it is not clear how they apply to health data retrieved through 'data.gov.sg,' especially in the case of dynamic datasets. Notably, the Singapore government uploaded 40 health datasets onto the portal two weeks after the AIHGle was published, perhaps reflecting its interest in leveraging public data to promote AI/ML innovation.

India

Background

India has the distinction of running the largest biometrics-driven digital identity program in the world, which has been operational since 2009.[70] The digital ID, called Aadhaar, is fashioned as a DPI used to authenticate the identity of Indian citizens seeking to avail of public and private services. Aadhaar may be considered as the first in a suite of DPIs that have since been implemented by the Indian government in sectors such as finance, logistics, and health. While some of these DPIs, such as DigiLocker – a cloud-based repository where an individual may choose to store electronic records pertaining to identity, educational and employment history, etc. – are designed as products, most middleware infrastructure built by the Indian state has taken the form of APIs and protocols. Examples include the Unified Payments Interface (a common railroad for instantaneous money transfers domestically), the Bharat Bill Payment System (an API-driven gateway for utilities payment),

the Goods and Services Tax Network (for collecting GST accrued to both the federal and local governments), etc.[71]

Since 2017, following its publication of a National Health Policy, the Indian government has sought also to incubate a "federated national health information architecture to roll out and link systems across public and private healthcare providers."[72] In 2018, India published a strategy paper on a National Health Stack (NHS), described as a "collection of cloud-based services" that run on open and interoperable APIs.[73] The strategy paper also mooted the creation of a digital health ID, a unique identifier that would allow Indian citizens to not only obtain longitudinal health records from a federated database, but also share it with healthcare providers anywhere in the country. Despite its ambitious goals, however, the NHS has struggled to materialize on account of two reasons. The domain of health is constitutionally the preserve of state governments in India, who have been reluctant to support a national initiative partly on account on lack of clarity on the implications of the technical infrastructure for their services.[74] Additionally, Indian regulators have also found it difficult to persuade large healthcare conglomerates to standardize and thereby render patient health records interoperable. As with Brazil and Singapore, however, the Indian government has attempted to use pandemic-era health surveillance powers to shift the momentum in its favor.

The role of APIs in the digitization of healthcare

To mitigate the spread of the coronavirus, India's National Health Authority (NHA) developed and mandated the use of two platforms for contact-tracing and COVID-19 vaccine management. Called Aarogya Setu ('Health Bridge') and CoWIN[75] respectively, the development and implementation of these applications provided the government with the institutional fillip needed to create a pan-Indian technical architecture for the NHS.[76] The NHA has sought to utilize not only the personnel resources it marshalled to develop these applications, but also the ties built with healthcare providers to coordinate CoWIN registrations and vaccine deliveries, towards the cause of the health stack. The NHS is envisioned as a "building block" comprising the following layers:[77]

a Data layer, which includes health IDs, longitudinal Personal Health Records (PHRs), registries of healthcare professionals and services, as well as other healthcare-related data such as hospital visit summaries, prescriptions, immunization records, etc.;
b A protocol layer, also known as the Unified Health Interface (UHI)[78] that comprises APIs enabling the seamless retrieval and sharing of data among various actors involved in the provision of healthcare services. Specifically, the UHI will comprise three categories of APIS: registry APIs, gateway APIs, and consent/information exchange APIs. Registry APIs facilitate the standardized collection and maintenance of data,

gateway APIs lay down specifications for access to particular healthcare networks, and consent APIs specify rules for "data fiduciaries," which are specialized entities that manage the consent of the user to share data with third parties;

c An application layer, featuring user-facing apps developed by public and private actors.

As the outline above indicates, the UHI is critical to smooth functioning of the NHS. The API-driven layer will not only determine who can access sensitive health data of Indian citizens, but also the granularity of the data so shared with different types of entities.

AI-enabled services and future plans

At the time of writing, the various policies and technical specifications that make up the NHS are in early stages of stakeholder consultations, but the proposed architecture of the NHS makes it clear APIs will invariably play an important role in ensuring access to training data for AI/ML services. Key to the use of health data for training ML algorithms will be the classification and labeling of data, which are part of the standardized PHRs. Additionally, a draft policy on data retention released alongside technical specifications refers to the conditions under which personal data may be anonymized or pseudonymized, as well as circumstances under which anonymized data should be deleted.[79] Nonetheless, the question remains as to the technical architecture that will facilitate the anonymization of data and concurrently, the use of training data in India's health sector. The NHA's blueprint for the health stack leaves this question open and suggests AI-enabled "clinical decision support systems" will be rolled out in Year 4 of its implementation.[80]

APIs, AI insecurities, and middle power diplomacy

The middleware architecture proposed or implemented by Brazil, Singapore, and India to digitize health data and render it available for training AI/ML models reveals the extent to which developing countries are reliant on APIs. Indeed, the three 'models' presented in this chapter are likely to be adopted by others states to jumpstart the development of AI-enabled services not only in health but also other sectors. States that have already made significant strides towards digitizing public datasets may opt, like Singapore, to make such data available via APIs but leave the selection of data and training of ML models to third parties. Those countries that have lagged behind in digitization may develop APIs to facilitate the standardized input of electronic records and the integration, subsequently, of national databases, as Brazil has done. In such cases, the respective entities responsible for digitizing health records may offer secondary APIs to facilitate third-party access. In yet other cases, states may not only standardize the creation of electronic records but

lay down strict policies and technical specifications – as India proposes – to determine how such data is shared with public and private actors alike.

All three approaches present security concerns for AI/ML services that may be exploited by strategic adversaries. Developing countries that make datasets available for third-party use may not have the regulatory capacity of a small, and relatively wealthy country like Singapore to monitor or enforce guidelines like the AIHGle. The use of APIs available on 'data.gov.sg' is governed by Terms of Service under Singapore's Open Data License, which not only restrict the use of such APIs to specific purposes but also prohibit downstream sub-licensing by third parties.[81] The Singapore model places a lot of trust in self-regulation by the market. For most emerging markets, however, a strong cybersecurity or data protection regulator is essential to monitor malicious ICT activity, because an infant private sector may have even lesser resources than the state to mitigate them. Cyber attacks by state or state-sponsored actors could specifically target API vulnerabilities to manipulate or destroy information in public databases, with a view to compromising AI-enabled services that rely on them. Additionally, poor API security on the part of private actors could have serious and adverse consequences for the integrity of the same training model data that is subsequently used by other developers for their respective services. Finally, the challenges of securing AI-enabled services grow in complexity when they rely on real-time APIs that provide 'live' data. At the time of writing, most datasets uploaded onto 'data.gov.sg' are static in nature. But as Singapore (and other countries) develop real-time APIs, regulators will need to find mechanisms that instantaneously identify and remedy serious cyber attacks on dynamic datasets, failing which they may cause lasting and widespread damage on AI-enabled services that rely on the same data. The risks associated with ML APIs and Man-in-the-Middle attacks have already been documented in this chapter, and they apply in particular to the Singapore model.

Brazil's API strategy is aimed at integrating electronic health records across the country, and making them available for governments at the federal, state, and local levels. As a result, AI/ML innovation in Brazil and other countries that follow its path may be more decentralized. Municipalities and local healthcare providers may tie up with research institutions and market actors to pilot AI-enabled services in provinces by granting them access to public data through their own APIs. The challenge inherent in this approach lies in testing and auditing the security of locally developed APIs that grant third-party access to national databases. Local governments may not spend as much time and resources reviewing their APIs for security flaws as a national regulator or agency. The cybersecurity of national infrastructure is only as strong as its weakest link. If security considerations are not adequately baked into the design lifecycle of such local APIs, states will be confronted with the same challenges identified above with respect to AI/ML services.

India's approach to standardizing electronic records and specifying rules for their sharing via APIs may seem tightly controlled, but presents its own set

of problems. Government control of API design and implementation, even in a democracy, can result in the API development process being opaque and unaccountable to outside stakeholders. Admittedly, this is a problem with the API-fication of public databases everywhere. However, India's wielding of executive power to force the adoption of DPI like Aadhaar[82] and Aarogya Setu, and the non-responsiveness of its bureaucracy to serious security incidents[83] raises the concern that a powerful government apparatus may be less receptive and agile to innovation. Additionally, with the state being the ultimate arbiter of key API decisions such as data labeling and the granularity of data sharing, its unaccountability vis-à-vis the research community and market actors can hamper investigations into security breaches of AI-enabled services.

The three countries whose plans for digital healthcare have been reviewed here not only stand out for contrasting API-led approaches to data sharing with third parties. They are also influential middle powers and democracies, whose successes in digitalizing their economies will be closely observed by regional and global actors alike. As they emulate attempts by these countries to open up public databases to third parties, states may, in the process, also replicate poor cybersecurity practices with respect to APIs.

Whatever the model, APIs are likely to emerge as vectors of cyber attacks on AI-enabled middleware services in middle powers. The COVID-19 pandemic has accelerated the digital transformation of their economies and societies, but attendant cyber risks have also risen. If the SingHealth system suffered a cyber attack in 2018, API-driven infrastructure in other middle powers, such as India's biometric ID system[84] and Brazil's Conecte-SUS platform[85] already suffered serious breaches during the pandemic. As governments rush to share data with the market through APIs for AI/ML innovation, they can also open a gateway for malicious actors. Handling as they do large volumes of information, APIs could be used to corrupt strategic information in databases about the demographic make-up of a country. Then there is the possibility of attack on the AI-enabled service itself. The damage caused by cyber attacks on training data and ML models will not only be economic, but political and psychological. Such attacks can erode trust in AI-enabled services among states and societies alike. In sensitive sectors such as health or transportation where real-time data is involved, cyber attacks can have catastrophic consequences, leading to human casualties.

How are middle powers likely to respond in geopolitical terms to this middleware dilemma?

The first possibility is that middle powers, including those states that have been traditionally reluctant to join alliances or plurilateral security arrangements, may seriously evaluate the possibility of collective measures to defend and even respond to cyber attacks on middleware infrastructure. Many middle powers are constrained by resources to build serious offensive and cyber capabilities. As a result, they have "over-invested" in publicly observable

efforts to build institutional cyber capacity (policies, regulatory agencies, etc.) with little deterrent effect to show for the same.[86] Were these states to open their databases and middleware infrastructure to AI-enabled services, whose security policies as well as algorithmic models are not always fully transparent, they would have even less control or oversight over their digital networks. To detect, prevent, and mitigate cyber attacks on critical resources, therefore, states may seek assistance from countries with more advanced capabilities. In particular, middle powers may enter into agreements with other states, including Great Powers, to protect their infrastructure from malicious cyber operations. A good example of collective cyber diplomacy in this regard is the Quad, a group consisting of the United States and three middle powers – Australia, India, and Japan. Motivated primarily by security compulsions in the wake of China's military "assertiveness" in Asia,[87] the Quad has committed to a number of cybersecurity initiatives, including a Quad Cybersecurity Partnership to share threat information about, develop software standards for, and build capacity to address cyber attacks on critical infrastructure.[88] Proposals for middle powers to develop collective diplomatic and military measures to mitigate major cyber threats are not new, and the 'middleware dilemma' identified in this chapter offers another compelling reason for such cooperation.[89] A second possibility is that middle powers could engage in cyber diplomacy to articulate and implement norms on the security of AI-enabled services. These norms, which may be articulated in intergovernmental or multistakeholder venues, may be comparable to guidelines on data and algorithmic risks identified by Singapore's AIHGLe or address AI vulnerabilities highlighted by other prominent regulators such as the European Union's ENISA.[90] States may also incubate or encourage market players to develop industry guidelines on API security, which has been lagging despite their growing importance to digital services, including AI-enabled services.

In summary, the 'middleware dilemma' will nudge middle powers to play a more active role in cybersecurity diplomacy, with a view to ensure the stability of cyberspace and to enhance their own capacities to address sophisticated cyber threats. Needing to sustain the digital transformation of key sectors, middle powers cannot afford to let discussions on AI security be shaped solely by Great Power politics: their proactive diplomacy could well lead to new norms or collective arrangements on the protection both of critical infrastructure as well as AI-enabled services that run on them.

Relevant disclosure and conflict of interest

The author is a volunteer with iSPIRT, a not-for-profit entity based in Bengaluru responsible for developing some API-driven digital public infrastructure for the Indian government. He is not involved in any technical or policy effort related to digital health data in the country, and as such, does not report any conflicts of interest.

Notes

1 Allan Patience, "Imagining middle powers," *Australian Journal of International Affairs* 68, no. 2 (15 March 2014): 214.
2 Eduard Jordaan, "The concept of a middle power in international relations: distinguishing between emerging and traditional middle powers," *Politikon* 30, no. 1 (May 2003): 167.
3 Charalampos Efstathopoulos, "Reinterpreting India's rise through the middle power prism," *Asian Journal of Political Science* 19, no. 1 (April 2011).
4 See generally, Jose Goldemberg, "Technological leapfrogging in the developing world science & technology," *Georgetown Journal of International Affairs* 12, no. 1 (2011).
5 Adam Chapnick, "The middle power," *Canadian Foreign Policy Journal* 7, no. 2 (January 1999): 73.
6 "Artificial intelligence index report 2021" (Stanford University Human-Centered Artificial Intelligence, November 2021), 155.
7 Johnny Kung, "Building an AI world: Report on national and regional AI strategies, second edition" (CIFAR, May 2020), 14.
8 Liv Marte Nordhaug and Kevin O'Neil, "Co-developing digital public infrastructure for an equitable recovery," *The Rockefeller Foundation* (blog), 22 July 2021.
9 Lalla Soundous Elkhaili El Alami, Asuka Nemoto, and Yoshinori Nakata, "Investigation of users' experiences for online access to their electronic health records in Japan," *Global Health & Medicine* 3, no. 1 (28 February 2021).
10 "PayNow Singapore," accessed 8 February 2022. https://abs.org.sg/consumer-banking/pay-now/.
11 "UPI: Unified payments interface – Instant mobile payments, *NPCI*," accessed 8 February 2022. https://www.npci.org.in/what-we-do/upi/product-overview.
12 "About your Aadhaar," *Unique Identification Authority of India, Government of India*, accessed 8 February 2022. https://uidai.gov.in/my-aadhaar/about-your-aadhaar.html.
13 Omidyar Network, "The open-source, identity platform MOSIP hits a new milestone," *Omidyar Network* (blog), 6 July 2020.
14 "ID-card," *e-Estonia*, accessed 8 February 2022. https://e-estonia.com/solutions/e-identity/id-card/.
15 See, "AI at the forefront of efforts to treat coronavirus patients," *GOV.UK*; "Big push for AI proves fruitful and useful," *TechNews Singapore Government*, 1 July 2020; Sarah O'Meara, "China's data-driven dream to overhaul health care," *Nature* 598, no. 7879 (6 October 2021): S1–3.
16 Mark Boyd, "Understanding what it takes to secure your API," *ProgrammableWeb*, 27 September 2017; "State of develops 2021" (Google Cloud, 2021), 24–25.
17 "Akamai: API attacks are exposing security vulnerabilities," *VentureBeat* (blog), 27 October 2021.
18 Tiffany Xingyu Wang and Matt McLarty, "APIs aren't just for tech companies," *Harvard Business Review*, 13 April 2021.
19 See generally, "API data breaches in 2020," *CloudVector* (blog), 23 December 2020.
20 Jason Macy, "API security: Whose job is it anyway?" *Network Security* 2018, no. 9 (1 September 2018): 6–9.
21 Deokyoon Ko, Kyeongwook Ma, Sooyong Park, Suntae Kim, Dongsun Kim, and Yves Le Traon, "API document quality for resolving deprecated APIs," in *2014 21st Asia-Pacific Software Engineering Conference* (2014), 27–30; "How zombie APIs pose a forgotten vulnerability," *Traceable App & API Security*, 28 May 2021.
22 Vickie Li, "API security 101: Excessive data exposure," *ShiftLeft*, 13 July 2021.

23 "OWASP Top 10:2021," accessed 8 February 2022. https://owasp.org/Top10/.
24 See generally, Chuan Zhao, Shengnan Zhao, Minghao Zhao, Zhenxiang Chen, Chong-Zhi Gao, Hongwei Li, and Yu-an Tan, "Secure multi-party computation: Theory, practice and applications," *Information Sciences* 476 (1 February 2019).
25 See generally, "MLOps with azure machine learning" (Microsoft); "Creating a machine learning-powered REST API with Amazon API gateway mapping templates and amazon sagemaker," *Amazon Web Services*, 13 March 2020.
26 Karl Manheim and Lyric Kaplan, "Artificial intelligence: Risks to privacy and democracy," *Yale Journal of Law and Technology* 21 (2019): 127–129.
27 Chengcheng Wan, Shicheng Liu, Henry Hoffmann, Michael Maire, and Shan Lu, "Are machine learning cloud APIs used correctly?" in *2021 IEEE/ACM 43rd International Conference on Software Engineering* (ICSE, 2021), 127.
28 Ibid., 128.
29 Ibid., 129.
30 Lingjiao Chen et al., "Did the model change? Efficiently assessing machine learning API shifts" (29 July 2021): 2.
31 Bruce Schneier and Trey Herr, "Russia's hacking success shows how vulnerable the cloud is," *Foreign Policy* (blog), 24 May 2021.
32 Andrey Krupskiy, Remmelt Blessinga, Jelmer Scholte, and Slinger Jansen, "Mobile software security threats in the software ecosystem, a call to arms," in *Software Business: 8th International Conference, ICSOB 2017* (Springer, 2017).
33 Katherine E. Bliss, "Brazil's Sistema Único Da Saúde (SUS): Caught in the cross fire," *Center for Strategic and International Studies*, 21 June 2017.
34 See generally, M. Sanni Ali, Maria Yury Ichihara, Luciana Cruz Lopes, George C.G. Barbosa, Robespierre Pita, Roberto Perez Carreiro, Djanilson Barbosa dos Santes, et al., "Administrative data linkage in Brazil: Potentials for health technology assessment," *Frontiers in Pharmacology* 10 (23 September 2019).
35 Ibid., 14.
36 Fernando Rocha Lucena Lopes, Karolinne Souza Monteiro, and Silvana Santos, "How data provided by the Brazilian information system of primary care have been used by researchers," *Health Informatics Journal* 26, no. 3 (1 September 2020).
37 Sanni Ali et al., "Administrative data linkage," 2.
38 Giliate Cardoso Coelho Neto, Rosemarie Andreazza, and Arthur Chioro, "Integração Entre Os Sistemas Nacionais de Informação Em Saúde: O Caso Do e-SUS Atenção Básica," *Revista de Saúde Pública* 55 (1 December 2021): 95.
39 "Conecte SUS pilot project: Final report in Alagos" (Brasilia: Ministry of Health, 2021), 3.
40 *Brazilian National Digital Health Strategy (2020–2028)* (Brasilia: Ministry of Health, 2020), 76–93.
41 "The new Brazilian strategy for artificial intelligence," *Offices of Science and Innovation*, accessed 8 February 2022. https://sweden-science-innovation.blog/brasilia/the-new-brazilian-strategy-for-artificial-intelligence/.
42 *Brazilian National Digital Health Strategy (2020–2028)*, 106.
431 *st Brazilian National Digital Health Strategy 2020–2028 Monitoring and Evaluation Report* (Brasilia: Ministry of Health, 2021), 18.
44 Ibid., 26.
45 "Portal de Serviços," accessed 8 February 2022. https://servicos-datasus.saude.gov.br/.
46 "Apple and Google's COVID-19 exposure notification API updated with improvements, Brazil launches app with alerts –9 to 5 Mac," *9to5mac.com*, 31 July 2020.
47 Olhar Digital, "Covid-19: Falha Na Plataforma e-SUS Gera Subnotificação de Casos No País," *Olhar Digital* (blog), 19 June 2020: 19. [translation by Safari].

48 Imprensa Nacional, "PORTARIA No 1.434, DE 28 DE MAIO DE 2020 – DOU – Imprensa Nacional," accessed 8 February 2022. https://www.in.gov.br/web/dou. [translation by Safari].
49 Imprensa Nacional, "PORTARIA GM/MS No 1.768, DE 30 DE JULHO DE 2021 – DOU – Imprensa Nacional," accessed 8 February 2022. https://www.in.gov.br/web/dou. [translation by Safari].
50 "The digitization of public services as a way to add value to the population," *Ideiagov*, accessed 8 February 2022. https://ideiagov.sp.gov.br/a-digitalizacao-dos-servicos-publicos-como-forma-de-agregar-valor-para-a-populacao/.
51 "Government tech for the people," *TechNews Singapore Government*, 23 May 2016.
52 "Our strategic national projects," accessed 8 February 2022. https://www.smartnation.gov.sg//initiatives/strategic-national-projects.
53 Irene Tham, "Personal info of 1.5 m SingHealth patients, including PM Lee, stolen in Singapore's worst cyber attack," *The Straits Times*, 20 July 2018.
54 "Digital Health," Health sciences authority (Singapore), accessed 8 February 2022. https://www.hsa.gov.sg/medical-devices/digital-health.
55 "*Regulatory Guidelines for Software Medical Devices – A Lifecycle Approach*" (Singapore: Health Sciences Authority, April 2020).
56 "How can Singapore's govtech stay number one?" *GovInsider* (blog), 16 January 16, 2020.
57 Wendell Santos, "How Singapore will run the country using APIs," *ProgrammableWeb*, 24 June 2018.
58 Ibid.
59 "Case study (APEX – Singapore)," in *Embracing Innovation in Government: Global Trends 2018* (OECD, 2018); "Inside Singapore's plans to share data across agencies," *GovInsider* (blog), 19 May 2017.
60 "API exchange (APEX) – A centralised data sharing platform for the public sector," *Singapore Government Developer Portal*.
61 "Data.Gov.Sg," *Data.gov.sg*, accessed 8 February 2022. https://data.gov.sg/.
62 Santos, "How Singapore will run."
63 "Health," *Data.gov.sg*, accessed 8 February 2022. https://data.gov.sg/group/health.
64 *Artificial Intelligence in Healthcare Guidelines (AIHGle)* (Singapore: Ministry of Health, October 2021).
65 Ibid., 5.
66 Ibid., 19.
67 Ibid., 38.
68 Ibid., 35.
69 Ibid., 12.
70 "What to know about aadhaar, India's biometric identity system," *Time*, 28 September 2018.
71 See generally, Vivek Raghavan, Sanjay Jain, and Pramod Varma, "India stack – digital infrastructure as public good," *Communications of the ACM* 62, no. 11 (November 2019).
72 *National Health Policy, 2017* (Government of India, 2017), 25.
73 *National Health Stack: Strategy and Approach* (Government of India, July 2018).
74 Smriti Mudgal Sharma, "National health stack: A job half well-done," *Ideas for India*, 10 September 2018.
75 CoWIN is an online portal, but Indian citizens can also register for vaccines on the portal through Aarogya Setu, the contact-tracing app. Mandatory registration through CoWIN was subsequently rolled back by the Indian government. "For 18+, On-site registration allowed at government vaccine centres," *NDTV.com*, 24 May 2021.

76 See Saurav Basu, "Effective contact tracing for COVID-19 using mobile phones: An ethical analysis of the mandatory use of the Aarogya Setu application in India," *Cambridge Quarterly of Healthcare Ethics* 30, no. 2 (2020).
77 See *Consultation Paper on Proposed Health Data Retention Policy* (National Health Authority, April 2021).
78 See *Consultation Paper on Unified Health Interface* (National Health Authority, March 2021).
79 *Consultation Paper on Proposed Health Data Retention Policy*, 30.
80 *National Digital Health Blueprint* (Government of India, 2017), 51.
81 See, "Singapore open data licence," *Data.gov.sg*, accessed 9 February 2022. https://data.gov.sg/open-data-licence.
82 Vindu Goel, "'Big brother' in India requires fingerprint scans for food, phones and finances," *The New York Times*, 7 April 2018, sec. Technology.
83 Aria Thaker, "In a year of data breaches, India's massive biometric programme finally found legitimacy," *Quartz*, 26 December 2018.
84 "Chinese hackers targeted aadhaar database, times group: Report," *NDTV.com*, 22 September 2021.
85 "Brazil health ministry website hit by hackers, vaccination data targeted," *Reuters*, 11 December 2021, sec. Technology.
86 See, Nadiya Kostyuk, "Deterrence in the cyber realm: Public versus private cyber capacity," *International Studies Quarterly* 65, no. 4 (17 December 2021).
87 "Quad: The China factor at the heart of the summit," *BBC News*, 24 May 2022, sec. India.
88 "*Quad Cybersecurity Partnership: Joint Principles*" (Government of Japan).
89 Lisa Davidson, "Analysing the characteristics of middle power cyber capability," in *European Conference on Cyber Warfare and Security* (Academic Conferences International Limited, 2017); Roland Paris, "*Can Middle Powers Save the Liberal World Order*?" (Chatham House, 2019); Louk Faesen, Tim Sweijs, Alexander Klimburg, and Giulia Tesauro, "The promises and perils of a minimum cyber deterrence posture: Considerations for small and middle powers" (The Hague Centre for Strategic Studies, April 2022); Sangbae Kim, "The inter-network politics of cyber security and middle power diplomacy: A Korean perspective" (East Asia Institute, 2014).
90 "Artificial intelligence cybersecurity challenges" (ENISA, 15 December 2020).

Bibliography

1st Brazilian National Digital Health Strategy 2020–2028 Monitoring and Evaluation Report. Brasilia: Ministry of Health, 2021. https://bvsms.saude.gov.br/bvs/publicacoes/1st_brazilian_national_digital_health_strategy.pdf.

"AI at the forefront of efforts to treat coronavirus patients," *GOV.UK,* 8 February 2022. https://www.gov.uk/government/news/ai-at-the-forefront-of-efforts-to-treat-coronavirus-patients.

"Akamai: API attacks are exposing security vulnerabilities." *VentureBeat* (blog), 27 October 2021. https://venturebeat.com/2021/10/27/akamai-apis-attacks-are-exposing-security-vulnerabilities/.

"API data breaches in 2020." *CloudVector* (blog), 23 December 2020. https://www.cloudvector.com/api-data-breaches-in-2020/.

"API exchange (APEX) – A centralised data sharing platform for the public sector." *Singapore Government Developer Portal,* 8 February 2022. https://www.developer.tech.gov.sg/technologies/data-and-apis/apex.

"Apple and Google's COVID-19 exposure notification API updated with improvements, Brazil launches app with alerts –9 to 5 Mac." *9to5mac.com*, 31 July 2020. https://9to5mac.com/2020/07/31/apple-and-googles-covid-19-exposure-notification-api-updated-with-improvements-brazil-launches-app-with-alerts/.

"Artificial intelligence cybersecurity challenges." *Report/Study ENISA*, 15 December 2020. Accessed 6 June 2022. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.

Artificial Intelligence in Healthcare Guidelines (AIHGle). Singapore: Ministry of Health, October 2021. https://www.moh.gov.sg/docs/librariesprovider5/eguides/1-0-artificial-in-healthcare-guidelines-(aihgle)_publishedoct21.pdf.

"Artificial intelligence index report 2021." *Stanford University Human-Centered Artificial Intelligence*, November 2021. Accessed 8 February 2022. https://aiindex.stanford.edu/wp-content/uploads/2021/11/2021-AI-Index-Report_Master.pdf.

Basu, Saurav. "Effective contact tracing for COVID-19 using mobile phones: An ethical analysis of the mandatory use of the Aarogya Setu application in India." *Cambridge Quarterly of Healthcare Ethics* 30, no. 2 (2020): 1–10. https://doi.org/10.1017/S0963180120000821.

"Big push for AI proves fruitful and useful." *TechNews Singapore Government*, 1 July 2020. https://www.tech.gov.sg/media/technews/big-push-for-ai-proves-fruitful-and-useful.

Bliss, Katherine E. "Brazil's Sistema Único Da Saúde (SUS): Caught in the cross fire." *Center for Strategic and International Studies*, 21 June 2017. Accessed 8 February 2022. https://www.csis.org/blogs/smart-global-health/brazils-sistema-unico-da-saude-sus-caught-cross-fire.

Boyd, Mark. "Understanding what it takes to secure your API." *ProgrammableWeb*, 27 September 2017. https://www.programmableweb.com/news/understanding-what-it-takes-to-secure-your-api/analysis/2017/09/27.

"Brazil health ministry website hit by hackers, vaccination data targeted." *Reuters*, 11 December 2021. https://www.reuters.com/technology/brazils-health-ministry-website-hit-by-hacker-attack-systems-down-2021-12-10/.

Brazilian National Digital Health Strategy (2020–2028). Brasilia: Ministry of Health, 2020.

"Case study (APEX – Singapore)." *Embracing Innovation in Government: Global Trends 2018*, OECD, 2018. https://www.oecd.org/gov/innovative-government/Singapore-case-study-UAE-report-2018.pdf.

Chapnick, Adam. "The middle power." *Canadian Foreign Policy Journal* 7, no. 2 (January 1999).

Chen, Lingjiao et al. "Did the model change? Efficiently assessing machine learning API shifts," arXiv:2107.14203 [stat.ML] (29 July 2021).

"Chinese hackers targeted Aadhaar database, times group: Report." *NDTV.com*, 22 September 2021. Accessed 6 June 2022. https://www.ndtv.com/india-news/chinese-hackers-targeted-aadhaar-database-times-group-report-2549166.

Consultation Paper on Proposed Health Data Retention Policy. New Delhi: National Health Authority, April 2021. https://abdm.gov.in/assets/uploads/consultation_papersDocs/Consultation_Paper_on_Health_Data_Retention_Policy_21.pdf.

Consultation Paper on Unified Health Interface. New Delhi: National Health Authority, March 2021. https://abdm.gov.in/assets/uploads/consultation_papersDocs/UHI_Consultation_Paper.pdf.

Conecte SUS Pilot Project: Final Report in Alagos. Brasilia: Ministry of Health, 2021. https://bvsms.saude.gov.br/bvs/publicacoes/conectesus_pilot_project_final_report.pdf.

"Creating a machine learning-powered REST API with amazon API gateway mapping templates and amazon sagemaker." *Amazon Web Services*, 13 March 2020. https://aws.amazon.com/blogs/machine-learning/creating-a-machine-learning-powered-rest-api-with-amazon-api-gateway-mapping-templates-and-amazon-sagemaker/.

Davidson, Lisa. "Analysing the characteristics of middle power cyber capability." In *European Conference on Cyber Warfare and Security*, 566–572, Reading: Academic Conferences International Limited, 2017. https://www.proquest.com/docview/1966799273/abstract/F35A1A01AE474C6FPQ/1.

Efstathopoulos, Charalampos. "Reinterpreting India's rise through the middle power prism." *Asian Journal of Political Science* 19, no. 1 (April 2011): 74–95. https://doi.org/10.1080/02185377.2011.568246.

El Alami, Lalla Soundous Elkhaili, Asuka Nemoto, and Yoshinori Nakata. "Investigation of users' experiences for online access to their electronic health records in Japan." *Global Health & Medicine* 3, no. 1 (28 February 2021): 37–43.

Faesen, Louk, Tim Sweijs, Alexander Klimburg, and Giulia Tesauro. "The promises and perils of a minimum cyber deterrence posture: Considerations for small and middle powers." *The Hague Centre for Strategic Studies*, April 2022. https://hcss.nl/report/promises-and-perils-of-minimum-cyber-deterrence-posture/.

"For 18+, on-site registration allowed at government vaccine centres." *NDTV.com*, 24 May 2021. Accessed 9 February 2022. https://www.ndtv.com/india-news/coronavirus-those-in-18-44-age-group-allowed-on-site-registration-appointment-on-cowin-for-vaccination-at-government-centres-2448313.

Goel, Vindu. "'Big brother' in India requires fingerprint scans for food, phones and finances." *The New York Times*, 7 April 2018. https://www.nytimes.com/2018/04/07/technology/india-id-aadhaar.html.

Goldemberg, Jose. "Technological leapfrogging in the developing world science & technology." *Georgetown Journal of International Affairs* 12, no. 1 (2011): 135–141.

"Government tech for the people." *TechNews Singapore Government*, 23 May 2016. https://www.tech.gov.sg/media/technews/government-tech-for-the-people.

"How can Singapore's Govtech stay number one?" *GovInsider* (blog), 16 January 2020. https://govinsider.asia/security/how-can-singapores-govtech-stay-number-one-chan-cheow-hoe-government-chief-digital-technology-officer/.

"How zombie APIs pose a forgotten vulnerability." *Traceable App & API Security*, 28 May 2021. https://www.traceable.ai/blog-post/how-zombie-apis-pose-a-forgotten-vulnerability.

"Inside Singapore's plans to share data across agencies." *GovInsider* (blog), 19 May 2017. https://govinsider.asia/innovation/api-exchange-apex-govtech-chan-cheow-hoe/.

Jordaan, Eduard. "The concept of a middle power in international relations: Distinguishing between emerging and traditional middle powers." *Politikon* 30, no. 1 (May 2003).

Kim, Sangbae. "*The Inter-Network Politics of Cyber Security and Middle Power Diplomacy: A Korean Perspective*." Seoul: East Asia Institute, 2014.

Ko, Deokyoon, Kyeongwook Ma, Sooyong Park, Suntae Kim, Dongsun Kim, and Yves Le Traon. "API document quality for resolving deprecated APIs." *21st Asia-Pacific Software Engineering Conference* 2 (2014): 27–30.

Kostyuk, Nadiya. "Deterrence in the cyber realm: Public versus private cyber capacity." *International Studies Quarterly* 65, no. 4 (17 December 2021): 1151–1162.

Krupskiy, Andrey, Remmelt Blessinga, Jelmer Scholte, and Slinger Jansen. "Mobile software security threats in the software ecosystem, a call to arms." in *Software Business: 8th International Conference, ICSOB 2017, Essen, Germany, June 12–13, 2017, Proceedings*, edited by Helena Holmström Olsson, Arto Ojala, and Karl Werder, pp. 161–175, Springer, 2017.

Kung, Johnny. "Building an AI world: Report on national and regional AI strategies." *CIFAR*, May 2020. https://cifar.ca/wp-content/uploads/2020/10/building-an-ai-world-second-edition.pdf.

Li, Vickie. "API security 101: Excessive data exposure." *ShiftLeft*, 13 July 2021. https://blog.shiftleft.io/api-security-101-excessive-data-exposure-a730d351fbae.

Lopes, Fernando Rocha Lucena, Karolinne Souza Monteiro, and Silvana Santos. "How data provided by the Brazilian information system of primary care have been used by researchers." *Health Informatics Journal* 26, no. 3 (1 September 2020): 1617–1630.

Macy, Jason. "API security: Whose job is it anyway?" *Network Security* 2018, no. 9 (1 September 2018): 6–9.

Manheim, Karl, and Lyric Kaplan. "Artificial intelligence: Risks to privacy and democracy." *Yale Journal of Law and Technology* 21 (2019): 127–129.

"MLOps with azure machine learning." *Microsoft,* 23 January 2022. https://azure.microsoft.com/en-us/resources/mlops-with-azureml/.

National Digital Health Blueprint. New Delhi: Ministry of Health and Family Welfare, Government of India, 2017. https://abdm.gov.in/home/ndhb.

National Health Policy, 2017. New Delhi: Ministry of Health and Family Welfare, Government of India, 2017. https://www.nhp.gov.in/nhpfiles/national_health_policy_2017.pdf.

National Health Stack: Strategy and Approach. New Delhi: National Institute for Transforming India (NITI Aayog), Government of India, July 2018. https://abdm.gov.in/publications/NHS_Strategy_and_Approach.

Neto, Giliate Cardoso Coelho, Rosemarie Andreazza, and Arthur Chioro. "Integração Entre Os Sistemas Nacionais de Informação Em Saúde: O Caso Do e-SUS Atenção Básica." *Revista de Saúde Pública* 55 (1 December 2021).

Nordhaug, Liv Marte, and Kevin O'Neil. "Co-developing digital public infrastructure for an equitable recovery." *The Rockefeller Foundation* (blog), 22 July 2021. https://www.rockefellerfoundation.org/blog/co-developing-digital-public-infrastructure-for-an-equitable-recovery/.

Olhar Digital. "Covid-19: Falha Na Plataforma e-SUS Gera Subnotificação de Casos No País." *Olhar Digital* (blog), 19 June 2020. https://olhardigital.com.br/2020/06/19/noticias/covid-19-falha-na-plataforma-e-sus-gera-subnotificacao-de-casos-no-pais/.

O'Meara, Sarah. "China's data-driven dream to overhaul health care." *Nature* 598, no. 7879 (6 October 2021): S1–S3. https://doi.org/10.1038/d41586-021-02694-1.

Omidyar Network. "The open-source, identity platform MOSIP hits a new milestone." *Omidyar Network* (blog), 6 July 2020. https://medium.com/omidyar-network/the-open-source-identity-platform-mosip-hits-a-new-milestone-ff9137610bed.

Paris, Roland. "*Can Middle Powers Save the Liberal World Order*?" Paris: Chatham House, 2019.

Patience, Allan. "Imagining middle powers." *Australian Journal of International Affairs* 68, no. 2 (25 March 2014): 210–224.

"Quad cybersecurity partnership: Joint principles." *Government of Japan,* 6 June 2022. https://www.mofa.go.jp/files/100348060.pdf.

"Quad: The China factor at the heart of the summit." *BBC News*, 24 May 2022. https://www.bbc.com/news/world-asia-india-61547082.

Raghavan, Vivek, Sanjay Jain, and Pramod Varma. "India stack – digital infrastructure as public good." *Communications of the ACM* 62, no. 11 (November 2019): 76–81.

"*Regulatory Guidelines for Software Medical Devices – A Lifecycle Approach*." Singapore: Health Sciences Authority, April 2020. https://www.hsa.gov.sg/docs/default-source/hprg-mdb/gudiance-documents-for-medical-devices/regulatory-guidelines-for-software-medical-devices---a-life-cycle-approach.pdf.

Sanni Ali, M., Maria Yury Ichihara, Luciana Cruz Lopes, George C.G. Barbosa, Robespierre Pita, Roberto Perez Carreiro, Djanilson Barbosa dos Santes et al. "Administrative data linkage in Brazil: Potentials for health technology assessment." *Frontiers in Pharmacology* 10 (23 September 2019): 1–20.

Santos, Wendell. "How Singapore will run the country using APIs." *ProgrammableWeb*, 24 June 2018. https://www.programmableweb.com/news/how-singapore-will-run-country-using-apis/else-where-web-case-study/2018/06/24.

Schneier, Bruce, and Trey Herr. "Russia's hacking success shows how vulnerable the cloud is." *Foreign Policy* (blog), 24 May 2021. https://foreignpolicy.com/2021/05/24/cybersecurity-cyberattack-russia-hackers-cloud-sunburst-microsoft-office-365-data-leak/.

Sharma, Smriti Mudgal. "National health stack: A job half well-done." *Ideas for India*, 10 September 2018. Accessed 9 February 2022. http://www.ideasforindia.in/topics/human-development/national-health-stack-a-job-half-well-done.html.

"State of develops 2021." *Google Cloud*, 2021. https://services.google.com/fh/files/misc/state-of-devops-2021.pdf.

Thaker, Aria. "In a year of data breaches, India's massive biometric programme finally found legitimacy." *Quartz*, 26 December 2018. Accessed 9 February 2022. https://qz.com/india/1501568/in-2018-supreme-court-backed-indias-aadhaar-despite-data-leaks/.

Tham, Irene. "Personal info of 1.5m SingHealth patients, including PM Lee, stolen in Singapore's worst cyber attack." *The Straits Times*, 20 July 2018. https://www.straitstimes.com/singapore/personal-info-of-15m-singhealth-patients-including-pm-lee-stolen-in-singapores-most.

Wan, Chengcheng, Shicheng Liu, Henry Hoffmann, Michael Maire, and Shan Lu. "Are machine learning cloud APIs used correctly?" *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, 2021. https://doi.org/10.1109/ICSE43902.2021.00024.

Wang, Tiffany Xingyu, and Matt McLarty. "APIs aren't just for tech companies." *Harvard Business Review*, 13 April 2021. https://hbr.org/2021/04/apis-arent-just-for-tech-companies.

"What to know about Aadhaar, India's biometric identity system." *Time*, 28 September 2018. Accessed 8 February 2022. https://time.com/5409604/india-aadhaar-supreme-court/.

Zhao, Chuan, Shengnan Zhao, Minghao Zhao, Zhenxiang Chen, Chong-Zhi Gao, Hongwei Li, and Yu-an Tan. "Secure multi-party computation: Theory, practice and applications." *Information Sciences* 476 (1 February 2019): 357–372.

6 Artificial intelligence and military superiority

How the 'cyber-AI offensive-defensive arms race' affects the US vision of the fully integrated battlefield

Jeppe T. Jacobsen and Tobias Liebetrau

In April 2016, then Deputy Secretary of Defense Robert O. Work presented the Department's latest strategic attempt to maintain US global military dominance, the *Third Offset Strategy*.[1] The strategic direction was clear: the American military must better integrate the latest data-driven, network-based and (semi)-autonomous technologies into its operative and organisational processes.[2] The strategy's underlying assumption is that the more data the US military can collect, process, analyse and share across the organisation, the better each operative unit will function on the battlefield. Artificial intelligence (AI) plays a vital role in this vision – not just as the technology that manages the increasing amount of sensor data collected from civilian and military systems but also as a tool to strengthen the resiliency of information systems, for example against cyberattacks.[3] However, it is still unclear whether AI is going to support or undermine the fully integrated, network-dependent battlefield of tomorrow.

So far, the emerging literature on the geopolitical and military implications of AI has paid attention primarily to how strategic stability is influenced by the international AI competition among great powers,[4] or to the specific military functions and processes that can be improved and undermined through the advancements in AI.[5] A main concern in the current literature is that the significant investments in AI in the United States, Russia, China and the EU – and their respective fears of being 'left behind' – will result in an arms race.[6]

This chapter makes a threefold contribution to the debate on a possible AI arms race. First, it argues that the risk of an AI arms race is not simply a matter of the undesirable framing by those political leaders and media outlets that do not understand the broader applications of AI.[7] Rather, the chapter shows that the AI arms race is closely linked to the military vision of technological superiority, which has dominated US military discourse since the *Second Offset Strategy* of the 1970s. Second, the chapter argues that the arms race that needs more attention does not relate simply to the general competition among great powers in the field AI, but relates also to the specific arms race between offensive and defensive cyber capabilities, which is currently

DOI: 10.4324/9781003284093-8

being accelerated by AI. The current attempts to improve cyber offense and cyber defence with AI hold both great potential and great risk to the ability for militaries to dominate the information space. Building on this finding, the chapter takes lesson from discussions on how militaries balance offense and defence in cyberspace, and finally argues that AI-enhanced cyber offensive capabilities are likely to dominate, which will have a negative impact on the likelihood that a fully integrated, networked battlefield will ever become reality.

The chapter comes in three parts. First, we situate AI in the broader US vision of military superiority to show that the idea of an AI arms race is deeply rooted in US military discourse. Second, having argued that AI-enhanced cyber capabilities will be both a prerequisite and a challenge to the US vision of a future battlefield, we take stock of the current state of AI in cyber offense and cyber defence. And third, we discuss three dynamics that have characterised the US approach to cyber offense and defence, namely the perception of offensive dominance, the role of intelligence agencies, and the emerging market for exploits. This is done to assess the implications of the on-going AI-cyber arms race for the possible realisation of a fully integrated future battlefield.

Imagining military superiority through technological 'offsets'

The heavy investment in AI among great powers has increasingly been described in the media as an 'AI arms race.'[8] Although the tendency to focus on killer robots borders hyperbole,[9] the characterisation is not without merit if 'arms race' is defined as competitive acquisition of military capability between two or more actors. Russia, China and the United States are all seeking to improve their military capability and efficiency by investing in AI innovation and they consider the competition a matter of who – in the famous words of President Putin – 'will become the ruler of the world.'[10] Yet, scholars have pointed out that the emphasis on an on-going 'AI arms race' risks becoming a self-fulfilling prophesy that prompts everyone to deploy unsafe AI-systems[11] and overlooks the many variations of technological proliferation and diffusion that must be managed very differently than if restraining an arms race.[12] While these points are both valid and important, the problem is not simply a 'framing problem' in the media, among policy makers, and in the military. To understand the importance of AI dominance in the US military today, it is necessary to return to the US first offset strategy of the early 1950s.

An offset strategy is an attempt to shift competitive focus in order to maintain an advantage over an emerging competitor despite facing restraints and disadvantages. President Eisenhower's 'New Look' strategy from the early 1950s represents the first example. Facing an intensifying competition with the Soviet Union and declining defence budgets after WWII, Eisenhower commissioned a review of the current defence policy, including an

exploration of alternative long term containment and deterrence strategies.[13] The result was a refocus on nuclear deterrence rather than the more expensive conventional deterrence.

After the Vietnam War, declining military expenses and increasing Warsaw Pact forces in Europe led US Defense Secretary Harold Brown to develop the *Second Offset Strategy*. The strategy saw new technologies as 'force multipliers' of combat effectiveness and as necessities for maintaining US military superiority. As a result, the Department of Defense funded a broad range of research and development initiatives on military-technological innovation, primarily within intelligence, surveillance, reconnaissance (ISR), precision guided weapons, stealth technologies, and space-based communication. While the 'offset' technologies were never tested against Soviet forces, they proved to be extremely efficient during the 1991 Operation Desert Storm. With the decisive US victory on the battlefields of Kuwait and Iraq, the Second Offset Strategy's fundamental assumption that (information) technological superiority leads to military superiority became the axiom of military thinking in the 1990s, particularly through the offshoot-concepts of a Revolution in Military Affairs (RMA) and Network Centric Warfare (NCW).[14]

Based on the rapid, private-sector-driven advancements of information technology in society in the late 1990s, NCW came to echo the second offset strategy but in the information age. Thus, information superiority was now the key to military superiority. NCW promised a transparent and predictable battlefield through the continuing development and integration of information technological innovations into the military to increase situational awareness, accelerate operational decision making, and get inside the adversary's decision circle to dictate the pace of military operations.[15] Yet, as Antoine Bousquet has convincingly shown,[16] NCW – although taking as its point of departure the idea of a new non-linear and more complex way of war – insisted simultaneously on pursuing a frictionless cybernetic war machine. Given this conceptual inconsistency as well as the inability to fulfil its promise of certainty, speed, and transparency in dealing with asymmetrical challenges in the war of Iraq and Afghanistan, the narrative of NCW lost some of its momentum.

However, rather than abandoning the underlying idea of information superiority as the key in warfighting, new concepts seemed only to reinforce the operational importance of dominating the information space. Cyber warfare – the ability to impact adversary computer networks for operational purposes – emerged as both a threat and opportunity in future wars. Relying on and adding to the already extensive cyber intelligence and data collection capacities in the National Security Agency, the DoD established the US Cyber Command in 2009 with the intent to defend DoD information networks as well as to direct and conduct full-spectrum military cyberspace operations.[17]

The rapid advances in AI and autonomy in the late 2010s further ignited the aspiration for even more certainty, speed, precision and coordination

across all domains. In fact, AI and autonomy became 'the technological sauce of the Third Offset Strategy' because of US inability to match 'tank for tank, plane for plane, person for person' of the resurgent Russia's and rising China's increasing ability to fight in all domains.[18] While also focusing on rethinking organisational and operational constructs, the Third Offset Strategy is, like its predecessor, primarily technological at its core. At an event at the Center for Strategic and International Studies in Washington D.C., Deputy Secretary Work underlined that injecting AI and autonomy into the C4I sensor grid of the US military would improve, for example, the ability to handle big data, determine patterns, provide timely relevant decision making through human-machine collaboration, and assist human operations through technology assistance such as wearables.[19]

The operational importance of AI and data as force multipliers supporting future multi-domain-operations is spelled out by the Army Capabilities Integration Center – Future Warfare Division in a 2018 White Paper. The White Paper concludes that 'artificial intelligence agents and algorithms will enable future force operations by processing, exploiting, and disseminating intelligence and targeting data.'[20] More recently, the National Security Commission on Artificial Intelligence in its final report accentuates the importance of AI for realising military superiority: 'our armed forces' competitive military-technical advantage could be lost within the next decade if they do not accelerate the adoption of AI across their missions.'[21]

The DoD has established the Joint Artificial Intelligence Center (JAIC) to ensure such acceleration in adopting AI across the US Defense. However, AI dominance does not only rely on the DoD. Since the Second Offset Strategy, the DoD perceives the ecosystem of collaboration between state, market and academia as the key driver of (military) technological innovation.[22] While the latest cyber-AI technologies developed by universities and profit-seeking private businesses show promising results, it comes with some risks as well as the last part of this section will show.

The inability to match the conventional forces of adversaries has therefore led the US defence apparatus to shift competitive focus to the latest technological developments in an attempt to maintain a military advantage. It is thus deeply ingrained in US military thinking that military superiority depends on technological superiority, which continues to rely on close collaboration and partnerships with the private sector. Hence, the current race to dominate within the field of AI is not simply a bad narrative or wrongful framing but part of a deep-rooted national security discourse where falling behind competitors constitute an existential threat.

As pointed out by Paul Scharre,[23] the danger of the AI arms race is that everyone rushes to deploy unsafe AI-systems, i.e. insufficiently tested AI weapon systems that can cause devastating unintended consequences. However, even thoroughly tested AI systems bear risks. The data-driven, networked, AI-infused, multi-domain US regime of warfighting imagined in the *Third Offset Strategy* comes indeed with its own paradox. The recent

attempt to achieve information superiority has resulted in a larger number of complex systems and system of systems. Such interconnected systems – while potentially improving military efficiency and capability – simultaneously broaden the attack surface and increase the vulnerability to cyberattacks.[24] Here, AI becomes a Janus-faced phenomenon as it promises to improve both network defence and network intrusion capabilities. The more pertinent AI arms race that will determine the future of the integrated battlefield is hence the one that currently takes place between offensive and defensive cyber capabilities. Before we go on to discuss how this arms race is likely to develop in the near term, it is necessary to briefly take stock of how AI is currently being developed and deployed to strengthen and streamline cyber defence and offensive techniques.

AI in cyber defence and offense – where are we now?

At its core, cyber defence is about knowing your network better than the adversary.[25] The fundamental challenge is, however, that an institution like the US Department of Defense has more than 15,000 networks and many million attached devices that need updates from time to time.[26] Such a task, when done manually, is time consuming and it is also near-impossible to make sure all systems work only as intended. It is, thus, no surprise that the research agency of the US Defense, DARPA, is heading efforts to automate software that identifies and patches vulnerabilities in IT-systems. Humans must be taken out of loop, as former head of NSA, General (ret.) Keith Alexander emphasised in a Senate hearing.[27]

An illustrative and much cited example of the attempt to take humans out of the loop of cyber defence was DARPA's 2016 Cyber Grand Challenge – the agency's attempt to get companies and universities to compete against each other to develop innovative AI solutions that could detect intrusions and identify, patch and exploit vulnerabilities. The winner system, *Mayhem* did not only win the first price of $2 million dollars, it also won a contract with the US Defense.[28] Mayhem was particularly innovative in its ability to balance performance and security when rolling out patches as well as in its ability to proactively make it more difficult to exploit vulnerabilities even before they are identified.[29] However, despite these improvements, Mayhem lost a similar 'capture-the-flag' competition at the hacker conference DEFCON against human-machine teams.[30] And while we have seen several example of AI systems consistently beating humans in games like Chess and Go or in military flight simulations exercises,[31] we still need to see successful examples of operative 'self-healing' IT-systems in use in the defence networks. In contrast, automated Intrusion Detection Systems have been implemented in the US Defense. The NSA-developed Sharkseer-programme uses different forms of AI techniques to conduct incoming traffic and e-mail inspections to identify zero-day exploits from the most advances threat actors.[32]

Yet, Sharkseer and the systems competing at the Cyber Grand Challenge are mainly built on rule-based AI and only to a lesser degree machine learning.[33] The publicly available cyber security solutions that are based on artificial neural networks and are trained through machine learning processes are currently mainly being developed in the private sector. The security company Fortinet, for example, has spent the last seven years developing a detection system based on supervised, unsupervised and reinforcement learning.[34] The hope is that the detection system will be able to identify and block malware in real time.

The potential of AI in cyber defence does not only relate to the identification and blocking of malware. A key global cyber security challenge is the reuse of code.[35] DARPA and US Air Force Research Laboratory are currently financing a project that seeks to develop a system that uses machine learning to evaluate the quality of source code and thereby help programmers reduce the risk of vulnerable software.[36] Some projects go even further. Rice University's Bayou-project – also financed by the US Defense – and projects at Cambridge University and Google are currently developing AI-systems that recognise and evaluate the intent of programmers, ultimately to avoid human typos and help streamline otherwise very complex coding.[37]

Despite several promising projects and large progress in the general development of AI, General Alexander's vision to get humans out of the cyber security loop still lies in the future. AI-systems still needs human assistance and few network administrators feel comfortable letting AI-systems independently re-programme code in critical, active environments, for example in the military. The 'Master AI' that runs the whole network, has access to billions of actions on a million devices and links data from all enterprise software applications and management systems, is still fiction.

Yet, it remains vital for the resiliency of the future integrated, network-based battlefield that an AI-enhanced cybersecurity system in milliseconds independently can quarantine suspicious activity, adapt firewalls and authentications, and isolate parts of the network.[38] Here, the more or less rational worry of network administrators are not the only obstacle. Machine learning algorithms need large quantities of data in a cloud where all data from all systems are stored in the right format – and not in a variety of formats across countless servers as is currently the case in most enterprises.[39] But even though, the technology to create a Master AI-system will most likely be here before we expect it, the will to actually let such a system run on an active critical network, knowing that mistakes and unintended consequences are inevitable and even part of the systems learning process, is probably not as imminent.

The reluctance is further fuelled by the fact that AI is not only relevant for cyber defence. As mentioned already, it has offensive potential as well. The company behind *Mayhem*, for example, has already developed and continues to experiment with different techniques aiming to undermine AI systems – either through identification and exploitation of vulnerabilities

in AI algorithms or through feeding adversary AI software with erroneous training data.[40] The rest of this section turns to the current development in AI-enhanced cyber offense.

The use of offensive cyber fuelled by AI is still at an early stage, but 'cyber attacks augmented by AI portend tailoring and manipulating the human side of important societal systems as well as introducing the risk that comes from moving technical skill from the hacker to an algorithm.'[41] At the same time, offensive and defensive cyber capabilities tend to develop together, as it is hard to safeguard the appropriate defence measures without knowledge of how state-of-the-art malicious cyber operations are executed.

AI and machine learning have already been used to improve the content of phishing e-mails, avoid spam-filters, and better map and systematise the collection of data on specific targets, which continues to be the foundation for most cyberattacks. The open source, neural network SNAP_R is an illustrative example of a very successful software that is able to map data across social media networks and send targeted phishing links.[42] Other current uses of AI to improve cyber offense relate to the attempt to determine the success rate of a cyberattack, to identify vulnerabilities in various software and networks, and to re-programme malware to avoid anti-virus.[43]

Most interestingly, AI can be used maliciously by integrating it with malware.[44] A myriad of options exist for malware developers to utilise AI. Simply put, malware will contain a definition of what purpose it is meant to serve and who it is intended for. The purpose may be to create two-way communication so that the attacker can copy data or encrypt files. Who the malware is intended for is also crucial. Should the malware simply compromise a specific user, or should it try to identify other target persons of the compromised user, for example by extracting data from contacts stored in the mail client, which can be used to escalate the attack. If purpose and goals are obscured, it is difficult to detect the malware.

At the Black Hat USA 2018 conference, IBM researchers presented 'a new breed of highly targeted and evasive attack tools powered by AI' called Deeplocker.[45] The malware 'conceals its intent until it reaches a specific victim' and 'it unleashes its malicious action as soon as the AI model identifies the target through indicators like facial recognition, geolocation and voice recognition.'[46] What made the Deeplocker malware extraordinary was the complex nature of the neural network that allowed it to conceal information about its target and the purpose deep in the code (hence the name). Moreover, Yu et al. underlines that 'even if the malware had been found, it would be very difficult for malware analysts to determine who or what the malware was searching for. Without knowing these things, it would be impossible to decipher the trigger condition, meaning that the payload would never be unlocked and remain unable to be studied.'[47]

Deeplocker and most of the other AI-infused techniques used for cyber offensive purposes are currently being developed in the private sector or as open source products. There is scarce amount of unclassified information on how

the DoD is current integrating AI into its offensive cyber capabilities in the US Cyber Command or the NSA. Yet, with the strategic focus in the DoD on maintaining technology superiority through collaboration with the private sector, it is very likely that the US military and intelligence agencies remain at the forefront of integrating AI into their cyber offensive capabilities.

This section has shown that AI provides great opportunities to revolutionise both defensive and offensive cyber capabilities. The development of preventive cyber AI – used to detect malware and prevent malicious cyber operations – is thus taking place in competition with the development of more sophisticated AI-integrated malware. The interactive symbiosis with escalating competition between the use of AI for defensive and offensive purposes can best be described as an AI arms race in cyberspace – a race that does not seem to stop for the time being.

But where are we heading? The question is vital as the United States is likely to continue its pursuit of its military vision of even more networked warfighting capabilities powered by integrated ICT as the key to military superiority. The next section takes lessons from the race between offensive and defensive cyber capabilities that has been on-going for more than two decades. If AI is simply tapping into and enhancing these existing dynamics, then the balance between offense and defence in cyberspace is instructive of where we are heading.

The cyber-AI arms race – lessons from balancing offense and defence in cyberspace

Jacquelyn Schneider has identified a capability/vulnerability paradox in modern digital-enabled warfare: networked technologies promise to improve military efficiency and capability but simultaneously broaden the attack surface, thus increasing the vulnerability to cyberattacks.[48] The trade-off between military efficiency and vulnerability mitigation depends on the balance between cyber defensive and offensive AI-enhanced capabilities. In other words, the cyber AI arms race is likely to determine whether or not the very efficient, fully integrated networked battlefield becomes reality. In the final part of the chapter, we introduce three dynamics that characterise the existing US approach to cyber offense and cyber defence: (1) the perception that cyber offense dominates cyber defence, (2) the dominant position of intelligence agencies in cyberspace, and (3) the reliance of the market for exploits. Each of the three is discussed with the view to how they are likely to influence the cyber-AI arms race and thus ultimately the US vision of fully integrated battlefield.

Cyber offensive dominance: technologically or socially determined?

Traditionally, we have experienced a consensus among US politicians and military that cyber security is dominated by the offence.[49] This has been

tied to the idea that cyberspace fundamentally changes international competition and warfare. As early as 1991, the National Academy of Sciences opened a report stating, 'We are at risk. Increasingly, America depends on computers. Tomorrow's terrorist may be able to do more damage with a keyboard than with a bomb.'[50] At the same time, the idea was born that readily-attainable cyber weapons would allow rouge states or individuals to cause massive destruction and the term electronic (later cyber) Pearl Harbor was coined. The endurance of both the idea and the term was depicted in 2012, when then-Secretary of Defense Leon Panetta became famous for saying that the United States faced the threat of cyber Pearl Harbor and a cyberattack perpetrated by nation states or violent extremists groups could be as destructive as the terrorists attack on 9/11.[51] At a scholarly and analytical level, the alleged systemic offensive advantage pertains to arguments regarding the design of the internet, endless amount of software vulnerabilities, the attribution problem and low barriers of entry.[52] Combined with the fact that securing networks are both difficult and incredibly resource intensive, then US Cyber Commander Michael Rogers argued that the United States must think about how to increase its capacity on the offensive side to get to that point of deterrence.[53]

In recent years, scholars have questioned the dominance of offensive cyber.[54] They offer different understandings of the role of technology and present different types of obstacles for the success of offensive cyber capabilities, including organisational, economic and knowledge/intelligence barriers. Rebecca Slayton argues, that the sources of cyber offensive or defensive advantage are not determined by technology alone. Instead of relying on technological determinism, she suggests to study the 'organizational processes that govern interactions between technology and skilled actors-processes such as software updating, vulnerability scanning, and access management.'[55] Slayton proposes that the apparent success of the offense stems from 'poor management and the relatively limited goals of offense, rather than a technologically *determine*d offensive advantage.'[56] Moreover, she asserts that one can only 'assess the offense-balance of cyber operations between two adversaries, but not of cyberspace,' because the balance is dyadic and not a systemic variable.[57]

Despite these scholarly insights, it has proven difficult to change the perception of offensive dominance in cyberspace among both practitioners and policy makers. As a result, US strategic documents continue to emphasise the insufficiency of relying on purely defensive measure when preventing and managing the cyberattacks.[58] In fact, the recent scholarly and political interest in the various ways in which cyberspace has reinforced a new form of strategic competition in the ambiguous space between war and peace has only made the US military's exploitation of vulnerabilities and hacking of foreign networks appear more relevant both for offensive and defensive purposes.[59] The US Cyber Commander Paul Nakasone emphasised that only by executing operations outside of US military networks – the so-called, defend

forward missions – is it possible for the United States to proactively defend and compete with adversaries in cyberspace.[60]

The perceived need to dominate the cyber offensive space will likely create incentive to integrate AI into these capabilities. Specifically, this means that the US military is likely to prioritise the development of AI-enhanced cyber capabilities that can improve the identification of software vulnerabilities, the targeting of cyberattacks, and the hiding of the payloads. However, the last decades of cyber capability development have taught us that capabilities are difficult to retain: Exploits are often dissected after they are used, the US military and intelligence community have not always been able to keep their capabilities secret, and the more capable adversaries have proven able to develop their own sophisticated capabilities. All these elements lead us back to the observed paradox that the US military goal of achieving information superiority through increased development and deployment of ICT, particularly AI, is likely to enhance the country's military capability, while simultaneously making it 'extremely vulnerable because of increasing dependencies on information.'[61] The increase in interconnected, ICT-enabled, and AI-enhanced systems – while promising to improve military efficiency and capability – simultaneously broadens the attack surface and increased the vulnerability to cyberattacks.[62]

The golden age of intelligence agencies: will they remain the most powerful in cyberspace?

While the debate about AI and cyber conflict often pivots on issues of military strategy and armed conflict, computers and networked systems have historically been built and used at the cutting edge primarily by the intelligence communities. The emergence of the Internet and networked computing thus provided intelligence agencies with an unprecedented opportunity for espionage and surveillance.[63] In the new digital environment, the same often flawed software and hardware have been used globally by governments, terrorists and criminals alike. When the amount of data exchanged simultaneously increased exponentially throughout the 1990s and 2000s, it is no surprise that intelligence agencies throughout the period continued to invest heavily in improving the collection capabilities in this new domain – and collected as much data as possible.[64] As a result, the intelligence agencies became the most capable players in cyberspace – and have been ever since. However, as a highly digitised society, the United States was deeply dependent on the same vulnerable IT-systems that the NSA and other US intelligence agencies were exploiting. As Ben Buchanan puts it, 'the means of secret stealing are in tension with the means of secret security.'[65]

NSA's initial answer to this paradox between collection and defence was the invention of the notion, NOBUS (Nobody But Us). Here, the premise was that the United States would seek to secure its networks and communication against all collection techniques except when these techniques

were deemed to be so complex that only the United States would be able to develop and use them.[66] The difficult judgments on other actors' cyber capabilities vis-à-vis American capabilities seemed to work initially. However, the Snowden revelations as well as cyber incidents such as Heartbleed and WannaCry increased the public demand for a more transparent process through which intelligence services and other public agencies judge whether an exploitable software vulnerability should be fixed for defence purpose or whether it should be retained for offensive or espionage purposes.

Until 2010 there was no process for sharing cyber threat intelligence between agencies or for working out the various equities between offensive and defensive mandates, but the US White House Cybersecurity Coordinators under President Obama and President Trump both responded to the public criticism by clarifying the official cross-institutional process through which such judgements were made. The process is known as the Vulnerabilities Equities Process (VEP).[67] If an agency wants to keep a zero day, it has to argue its case through the VEP to an Equities Review Board chaired by the National Security Council (NSC) and attended by representatives from other public agencies, including those most concerned with the security of critical U.S. infrastructure like the Department of Homeland Security (DHS) and the Department of Commerce.

Following the 2017 global cyber incidents WannaCry and NotPetya, the VEP made headlines. One of the reasons why WannaCry and NotPetya spread so quickly, globally and massively was because attackers used an exploit developed by – and later stolen from – the NSA (known as EternalBlue). Consequently, the US government and the NSA were criticised for their continued practice of stockpiling vulnerabilities for later use and leaving citizen and industry users vulnerable.[68] Moreover, critics have emphasised that even after the 2017 improvement to the process, it still contains open-ended language of the 'exception to disclosure.'[69] In other words, the decision on when to disclose and when to retrain IT vulnerabilities that can be used offensively still depends on the internal power hierarchy among the entities working with cybersecurity, cyber espionage and cyber warfare in the United States – a hierarchy that has NSA at the top. This has led Knake and Schwartz to recommend a transfer of the role of VEP executive secretary functions from the NSA to the more defensively minded DHS, an increase in transparency through an annual public report and a strengthened independent oversight.[70]

As these recommendations have not been implemented, the NSA's preference for intelligence collection continues to dominate – even if sometimes at the expense of individual cyber security.[71] The agency is therefore likely to continue its exploration of the various ways in which to improve its exploitation techniques with AI. The US adoption of the cyber doctrine of persistent engagement and defend forward will only strengthen the need for intelligence agencies. The NSA's development of AI-enabled cyber capabilities that are even more effective in identifying and exploiting IT vulnerabilities will then reinforce the long-standing debate on whether to retain zero day

vulnerabilities for intelligence purposes, use them in military operations or disclose them to vendors so they can be patched. In short, if the institutional power dynamics in the United States remain unchanged, then the dominant norm in cyberspace is likely to be an intelligence norm, where intelligence agencies seek to integrate AI into their exploitation capabilities. This would jeopardise the US military goal of achieving information superiority through increased development and deployment of AI.

The market for exploits: can the private sector be tamed?

ICT technologies and infrastructures are primarily developed, owned, operated and controlled by private companies and cyberspace is often depicted as a non-state-centric environment.[72] Unsurprisingly then, as computers became the primary tool and target of espionage and crime during the 2000s, a market developed around cyber security services. A part of this market relied on individual hackers searching for vulnerabilities and selling them to software companies or specialised cyber security companies. The latter could then use the vulnerabilities to simulate a cyberattack on a system, and ultimately help the companies in improving their intrusion detection software. However, as demonstrated by Nicole Perlroth these 'defence' dynamic was outcompeted by government agencies who could pay more.[73] Perlroth tells the story of the cyber security company iDefense that in the early 2000s started to receive calls from various government entities willing to pay $150,000 for a bug iDefense was buying from individual hackers for $400.[74]

The VEP described above is thus only a small part of a far larger ecosystem of vulnerability and disclosure. According to Jason Healey,[75] the ecosystem 'includes security researchers who find new vulnerabilities, vendors who patch them and perhaps seek them out through a corporate or independent "bug bounty" program, grey markets and other intermediaries who help broker connecting researchers to vendors (to patch) or attackers (to gain illicit entry), and government agencies that are sometimes attackers and sometimes defenders.' Today, governmental agencies regularly acquire cyber tools, including a wide range of cyber weapons, from private actors. The zero-days markets trade in vulnerabilities and exploits, 'the relative proportion of which differ according to whether a market is characterised as white, grey or black.'[76] US government agencies are said to be the main purchasers of vulnerabilities and exploits on the grey market.[77] One of the most discussed examples is the 2016 unlocking of the iPhone used by a terrorist in the San Bernardino shooting. The phone was unlocked by a small Australian hacking firm, ending a historic standoff between the U.S. government and Apple.[78]

As of 2021, there is no international agreement on if or how cyber weapons should/can be regulated, nor does any such regime seem to be pending. According to Tim Stevens,[79] the lack of a global governance regime for cyber weapons can be explained by the constraining power of the Tallinn Manual, US involvement in markets for cyber weapons, the institutional power of

internet technologies, and diplomatic claims to sovereignty that mask operations of compulsory power. Yet, there are attempts to build international import-export control agreements. The most famous – The Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies – is a multilateral arms control treaty between 42 countries that has increasingly seen the addition of software to the list of prohibited trade items. Initially focused on traditional weapons, war and unique products like chemicals, the Wassenaar Agreement has been updated in recent years to include software that encompasses intrusion command-and-control features.

However, the US and private tech companies worked actively to limit the scope of the controls, as they feared overly-broad controls limiting researchers ability to identify and correct security vulnerabilities and criminalising essential tools for stopping malware. Without further and firm support of the United States, it is unlikely that voluntary international regulation mechanisms like the Wassenaar Agreement will achieve its desired effect and proper restrictions be put in place. On the contrary, Stevens argues that the US' position as 'the dominant producer and consumer of cyber weapon components and research disincentives market regulation and encourages international trade in code entities like zero-day exploits.'[80] A global governance framework for cyber weapons is only emerging hesitantly. As the development of an effective architecture for regulating and prohibiting weapons is generally a slow process,[81] the prospect of governance of AI fuelled offensive cyber capabilities is discouraging.

In sum, the continued domination of intelligence thinking, the grey market trading by government agencies purchasing cyber weapon components, and the lack of import-export restrictions on cyber weapons all suggest that the cyber-arms race is likely to continue – also with the addition of AI. This risk further undermining the long-term stability and security of/in cyberspace for numerous reasons. It increases the general risk of accidents and incidents, it creates a strong incentive for research to be aimed at writing exploits rather than detecting and reporting vulnerabilities, it does not encourage software vendors to internalise their security costs, it crowds out defensive, resilience and stability efforts stimulated by other government agencies, and it contributes further to the strategic predicament of an inevitable cyber AI arms race.

Conclusion: the cyber AI-arms race is deep rooted but not inevitable

The chapter paints a bleak picture: The current race to dominate within the military field of AI is not simply a bad narrative or wrongful framing that can be easily replaced but is part of a deep-rooted national security discourse where falling behind competitors constitute an existential threat. The chapter also pointed to several reasons why investment in AI for

cyber-offensive purposes is likely to continue and dominate cyber-defence AI. Such offensive-skewed cyber-AI arms race risks undermining the US vision of military superiority through a fully-integrated, networked battlefield: Either adversaries develop AI-infused tools to undermine the integrity and accessibility of US battle networks or commanders and decision makers develop a lack of trust in the operability of these networks. As a result, AI could lose its appeal as the object that promises military superiority, and a new technology will emerge that promises battlefield dominance when fully developed and integrated sometimes in the future.

A conclusion about an inevitable arms race, however, is not written in stone. Both the scholarly and practical debates on cyber offense and cyber defence in the United States offer alternative paths. Slayton has already shown how particular organisational and management cultures are reproducing cyber offense as the dominant perspective, rather than being technologically determined.[82] Knake and Schwartz have introduced a range of policy recommendations that seek to strengthen the role of cyber defensive and cybersecurity apparatus in the United States vis-à-vis the intelligence agencies and the US Cyber Command.[83] Lastly, the attempts to promote a cyber export control regime – inspired e.g. by the Wassenaar Agreement – seek to weaken a private market for exploits that reproduce offensive dominance in the cyber domain. Each of these debates, if translated into policy, would strengthen the private businesses that work to strengthen the defensive side of the offensive-defensive arms race through new innovative AI solutions.

Notes

1 Robert O. Work, "Remarks by deputy secretary work on third offset strategy," *Speech, Brussels, Belgium*, 28 April 2016.
2 Jesse Ellman, Lisa Samp, and Gabriel Coll, "Assessing the Third Offset Strategy" (Center for Strategic & International Studies, 2017).
3 U.S. DoD, "Summary of the 2018 department of defense artificial intelligence strategy – harnessing AI to advance our security and prosperity" (U.S. Department of Defense, 2018).
4 Michael C. Horowitz, "Artificial intelligence, international competition, and the balance of power," *Texas National Security Review* 1, no. 3 (2018); Michael C. Horowitz et al., "Artificial Intelligence and International Security" (Center for New American Security, 2018); James Johnson, "Artificial intelligence & future warfare: Implications for international security," *Defense & Security Analysis* 35, no. 2 (2019); James Johnson, "Artificial intelligence: A threat to strategic stability," *Strategic Studies Quarterly* 14, no. 1 (2020).
5 Kareem Ayoub and Kenneth Payne, "Strategy in the age of artificial intelligence," *Journal of Strategic Studies* 39, no. 5–6 (2016); Benjamin M. Jensen, Christopher Whyte, and Scott Cuomo, "Algorithms at war: The promise, peril, and limits of artificial intelligence," *International Studies Review* 25 (June 2019); Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (W.W. Norton & Company, 2018).
6 Edward Moore Geist, "It's already too late to stop the AI arms race—we must manage it instead," *Bulletin of the Atomic Scientists* 72, no. 5 (2016); Michael T. Klare, "AI arms race gains speed," *Arms Control Today* 49, no. 2 (2019).

7 Cf. Heather M. Roff, "The frame problem: The AI "arms race" isn't one," *Bulletin of the Atomic Scientists* 75, no. 3 (2019); Paul Scharre, "Killer apps: The real dangers of an AI arms race," *Foreign Affairs* 98, no. 2 (2019).
8 Tom Simonite, "For superpowers, artificial intelligence fuels new global arms race," *Wired*, 8 September 2017; Julie E. Barnes and Josh Chin, "The new arms race in AI," *The Wall Street Journal*, 2 March 2018; Peter Apps, "Commentary: Are China, Russia winning the AI arms race?" *Reuters*, 15 January 2019.
9 Melissa K. Chan, "Could China develop killer robots in the near future? Experts fear so," *Time*, 13 September 2019; Matt Bartlett, "The AI arms race in 2019," *Towards Data Science* (blog), 28 May 2019.
10 Horowitz, "Artificial intelligence, international competition;" Elsa B. Kania, "Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power" (Center for New American Security, 2017).
11 Scharre, "Killer apps."
12 Roff, "The frame problem."
13 Robert Martinage, "Towards a New Offset Strategy – Exploiting U.S. Long-Term Advantages to Restore U.S. Global Power Projection Capability" (Center for Strategic and Budgetary Assessments, 2014), 6–13.
14 Alvin Toffler and Heidi Toffler, *War and Anti-War: Survival at the Dawn of the 21st Century* (Warner Books, 1994); David S. Alberts, John Garstka, and Frederick P. Stein. *Network Centric Warfare: The Face of Battle in the 21st Century* (National Defense University Press, 1999); William A. Owens and Edward Offley, *Lifting the Fog of War* (Johns Hopkins University Press, 2001); Andrew F. Krepinevich, "The Military-Technical Revolution: A Preliminary Assessment" (Center for Strategic and Budgetary Assessments, 2002).
15 Department of Defense, "Network Centric Warfare – Creating a Decisive Warfighting Advantage" (Department of Defense, 2003), 3; Prem Chand, "Network-centric warfare – some fundamentals," *Air Power* 2, no. 1 (2005).
16 Antoine Bousquet, *The Scientific Way of Warfare: Order and Chaos on the Battlefields of Modernity* (Columbia University Press, 2009), 185–234.
17 USCYBERCOM, "U.S. cyber command factsheet" (16 April 2014).
18 Work, "Remarks."
19 Cheryl Pellerin, "Deputy secretary: Third offset strategy bolsters America's military deterrence," *U.S. Department of Defence*, 31 October 2016.
20 Army Capabilities Integration Center – Future Warfare Division, *Operationalizing Robotic and Autonomous Systems in Support of Multi-Domain Operations*, White paper, 30 November 2018: 35.
21 National Security Commission on Artificial Intelligence, *Final Report* (2021), 9.
22 Donald J. Trump, "Executive order on maintaining American leadership in artificial intelligence," *The White House*, 2019; The White House, Office of Science and Technology Policy, "Artificial intelligence initiative: Year one annual report" (2020).
23 Scharre, "Killer apps."
24 Jacquelyn Schneider, "Digitally-Enabled Warfare: The Capability-Vulnerability Paradox" (Center for New American Security, 2016); Jacquelyn Schneider, "The capability/vulnerability paradox and military revolutions: Implications for computing, cyber, and the onset of war," *Journal of Strategic Studies* 42, no. 6 (2019); Keith F. Joiner and Malcom G. Tutty, "A tale of two allied defence departments: New assurance initiatives for managing increasing system complexity, interconnectedness and vulnerability," *Australian Journal of Multi-Disciplinary Engineering* 14, no. 1 (2018); Jennifer McCardle, "Victory Over and Across Domains: Training for Tomorrow's Battlefield" (Center for Strategic and Budgetary Assessments, 2019).
25 Rob Joyce, "Disrupting nation state hackers," presented at the USENIX Enigma, San Francisco, CA, 27 January 2016.

26 Scharre, *Army of None*, 216.
27 Ibid.
28 Thanassis Avgerinos et al., "The MAYHEM CYBER REASONING SYSTEM," *IEEE Security & Privacy* 16, no. 2 (2018).
29 Ibid.
30 Bruce Schneier, "The Coming of AI Hackers" (Belfer Center for Science and International Affairs, 2021), 22.
31 Ayoub and Payne, "Strategy in the Age," 803–804; Kania, "Battlefield Singularity," 27.
32 Satoru Mori, "US defense innovation and artificial intelligence," *Asia-Pacific Review* 25, no. 2 (2018): 29.
33 Horowitz et al., "Artificial Intelligence and International Security," 3. Rule-based AI is based on predefined facts about different situations and predefined rules about how these should be dealt with. In machine learning the algorithm is trained based on the outcome of a situation.
34 John Maddison, "Using advanced AI to stay ahead of cybercriminals," *Infradata* (blog), 29 August 2019.
35 Michael Sulmeyer and Kathryn Dura, "Beyond killer robots: How artificial intelligence can improve resilience in cyber space," *War on the Rocks* (blog), 6 September 2018.
36 Art Jahnke, "Could deepcode AI make life harder for hackers?" *Boston University – The Brink*, 11 December 2018.
37 Richard A. Clarke and Robert K. Knake, *The Fifth Domain: Defending Our Country, Our Companies, and Ourselves in the Age of Cyber Threats* (Penguin Press, 2019), 80–81.
38 Ibid., 246–247.
39 Ibid.
40 Scharre, *Army of None*, 221–222.
41 Christopher Whyte, "Poison, persistence, and cascade effects," *Strategic Studies Quarterly* 14, no. 4 (Winter 2020): 18.
42 John Seymour and Philip Tully, "Weaponizing Data Science for Social Engineering: Automated E2E Spear Phishing on Twitter," Conference paper, Black Hat USA 2016: 1–3.
43 Kelly Sheridan, "How attackers use machine learning to predict BEC success," *DARKReading*, 26 July 2017; Hyrum S. Anderson et al., "Evading Machine Learning Malware Detection," Conference paper, Black Hat USA 2017: 1, 3–4.
44 Malware is malicious software that pierces and damages a computer system without consent and without informing the system owner. The term is used to express a variety of forms of software or program code, such as computer virus, worm, Trojan horse and botnet.
45 M.P. Stoecklin, J. Jang, and D. Kirat, "Deeplocker: How AI can power a stealthy new breed of malware," *Security Intelligence*, 8 August 2018.
46 Ibid.
47 Ning Yu et al., "AI-powered GUI attack and its defensive methods," *Proceedings of the 2020 ACM Southeast Conference*, Session 1 – Full papers, 2–4 April 2020: 2.
48 Schneider, "Digitally-enabled warfare," 4.
49 Erik Gartzke and Jon R. Lindsay, "Weaving tangled webs: Offense, defense, and deception in cyberspace," *Security Studies* 24, no. 2 (2015); Rebecca Slayton, "What is the cyber offense-defense balance? Conceptions, causes, and assessment," *International Security* 41, no. 3 (2017).
50 National Academy of Science, *Computers at Risk: Safe Computing in the Information Age* (The National Academies Press, 1991), 7.
51 Jon R. Lindsay, "Stuxnet and the limits of cyber warfare," *Security Studies* 22, no. 3 (2013).

52 Joseph S. Nye, "Cyber Power" (Belfer Center for Science and International Affairs, 2010); Ben Buchanan, *The Cybersecurity Dilemma – Hacking, Trust, and Fear Between Nations* (Oxford University Press, 2017).
53 US State Senate. *Hearing to Receive Testimony on U.S. Strategic Command, U.S. Transportation Command, and U.S. Cyber Command in Review of the Defense Authorization Request for Fiscal Year 2016 and the Future Years Defense Program* (US State Senate, 2015).
54 Thomas Rid, *Cyber War Will Not Take Place* (Oxford University Press, 2013); Gartzke and Lindsay, "Weaving Tangled Webs;" Slayton, "What is the cyber offense-defense."
55 Slayton, "What is the cyber offense-defense," 74.
56 Ibid., 75.
57 Ibid., 74.
58 USCYBERCOM, "Achieve and maintain cyberspace superiority." *Command Vision for US Cyber Command* (2018); Donald J. Trump, "National cyber strategy of the United States of America" (The White House, 2018).
59 Lucas Kello, *The Virtual Weapon and International Order* (Yale University Press, 2017); Michael Fischerkeller and Richard J. Harknett, "Deterrence is not a credible strategy for cyberspace," *Orbis* 61, no. 3 (2017); Jason Healey, "The implications of persistent (and permanent) engagement in cyberspace," *Journal of Cybersecurity* 5, no. 1 (2019); Richard J. Harknett and Max Smeets, "Cyber campaigns and strategic outcomes," *Journal of Strategic Studies* (March 2020).
60 Paul M. Nakasone and Michael Sulmeyer, "How to compete in cyberspace," *Foreign Affairs*, 25 August 2020.
61 Schneider, "The capability/vulnerability paradox," 842.
62 Schneider, "Digitally-Enabled Warfare;" "The capability/vulnerability paradox;" Joiner and Tutty, "A tale of two allied defence departments;" McCardle, "Victory over and across."
63 Christopher Whyte and Brian Mazanec, *Understanding Cyber Warfare: Politics, Policy and Strategy* (Routledge, 2019): Chapters 4 and 5.
64 Ellen Nakashima and Joby Warrick, "For NSA chief, terrorist threat drives passion to "collect it all,"" *Washington Post*, 14 July 2013.
65 Ben Buchanan, "The Rise and Fall of the Golden Age of Signals Intelligence," Aegis Series Paper No. 1708, A Hoover Institution Essay, 2017: 2.
66 Ibid.
67 Michael Daniel, "Heartbleed: Understanding when we disclose cyber vulnerabilities," *White House Blog*, 28 April 2014; Rob Joyce, "Improving and making the vulnerability equities process transparent is the right thing to do," *Whitehouse.Gov* (blog), 15 November 2017.
68 K. Kristoffer Christensen and Tobias Liebetrau, "A new role for 'the public'? Exploring cyber security controversies in the case of WannaCry," *Intelligence and National Security* 34, no. 3 (2019).
69 Mimansa Ambastha, "Taking a hard look at the vulnerabilities equities process and its national security implications," *Berkeley Technology Law Journal* (blog), 23 April 2019.
70 Robert Knake and Ari Schwartz, "Government's role in vulnerability disclosure: Creating a permanent and accountable vulnerability equities process" (June 2016).
71 Jeppe T. Jacobsen, "Lacan in the US cyber defence: Between public discourse and transgressive practice," *Review of International Studies* 46, no. 5 (2020): 718–719.
72 Alex S. Wilner, "US cyber deterrence: Practice guiding theory," *Journal of Strategic Studies* 43, no. 2 (2020).
73 Nicole Perlroth, *This Is How They Tell Me the World Ends: The Cyberweapons Arms Race* (Bloomsbury Publishing, 2021).

74 Perlroth, *This Is How They Tell Me*, 39.
75 Jason Healey, "The US government and zero-day vulnerabilities: From pre-heartbleed to shadow brokers," *Columbia Journal of International Affairs* (November 2016): 3.
76 Tim Stevens, "Cyberweapons: Power and the governance of the invisible," *International Politics* 55 (2018): 489.
77 Ibid., 490.
78 Ellen Nakashima and Reed Albergotti, "The FBI wanted to unlock the San Bernardino shooter's iPhone. It turned to a little-known Australian firm," *Washington Post*, 14 April 2021.
79 Stevens, "Cyberweapons."
80 Ibid., 491.
81 Brian M. Mazanec, *The Evolution of Cyber War: International Norm for Emerging-Technology Weapons* (Potomac Books/University of Nebraska Press, 2015).
82 Slayton, "What is the cyber offense-defense."
83 Knake and Schwartz, "Government's role."

Bibliography

Alberts, David S., John Garstka, and Frederick P. Stein. *Network Centric Warfare: The Face of Battle in the 21st Century*. Washington, DC: National Defense University Press, 1999.

Ambastha, Mimansa. "Taking a hard look at the vulnerabilities equities process and its national security implications." *Berkeley Technology Law Journal* (blog), 23 April 2019. https://btlj.org/2019/04/taking-a-hard-look-at-the-vulnerable-equities-process-in-national-security/.

Anderson, Hyrum S., Anant Kharkar, Bobby Filar, and Phil Roth. "Evading machine learning malware detection." *Conference Paper, Black Hat USA 2017*, 22–27 July 2017, Las Vegas, NV, USA. https://www.blackhat.com/docs/us-17/thursday/us-17-Anderson-Bot-Vs-Bot-Evading-Machine-Learning-Malware-Detection-wp.pdf.

Apps, Peter. "Commentary: Are China, Russia winning the AI arms race?" *Reuters*, 15 January 2019. https://www.reuters.com/article/us-apps-ai-commentary-idUSKCN1P91NM.

Army Capabilities Integration Center – Future Warfare Division. *Operationalizing Robotic and Autonomous Systems in Support of Multi-Domain Operations*. White paper, 30 November 2018.

Avgerinos, Thanassis, David Brumley, John Davis, Ryan Goulden, Tyler Nighswander, Alex Rebert, and Ned Williamson. "The mayhem cyber reasoning system." *IEEE Security & Privacy* 16, no. 2 (2018): 52–60.

Ayoub, Kareem, and Kenneth Payne. "Strategy in the age of artificial intelligence." *Journal of Strategic Studies* 39, no. 5–6 (2016): 793–819.

Barnes, Julie E., and Josh Chin. "The new arms race in AI." *The Wall Street Journal*, 2 March 2018. https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261.

Bartlett, Matt. "The AI arms race in 2019." *Towards Data Science* (blog), 28 May 2019. https://towardsdatascience.com/the-ai-arms-race-in-2019-fdca07a086a7.

Bousquet, Antoine. *The Scientific Way of Warfare: Order and Chaos on the Battlefields of Modernity*. New York, NY: Columbia University Press, 2009.

Buchanan, Ben. *The Cybersecurity Dilemma – Hacking, Trust, and Fear Between Nations*. New York: Oxford University Press, 2017.

Buchanan, Ben. *The Rise and Fall of the Golden Age of Signals Intelligence*. Aegis: Hoover Institution, 2017.

Chan, Melissa K. "Could China develop killer robots in the near future? Experts fear so." *Time*, 13 September 2019. https://time.com/5673240/china-killer-robots-weapons/.

Chand, Prem. "Network-centric warfare – some fundamentals." *Air Power* 2, no. 1 (2005): 1–24.

Christensen, K. Kristoffer, and Tobias Liebetrau. "A new role for 'the public'? Exploring cyber security controversies in the case of WannaCry." *Intelligence and National Security* 34, no. 3 (2019): 395–408.

Clarke, Richard A., and Robert K. Knake. *The Fifth Domain: Defending Our Country, Our Companies, and Ourselves in the Age of Cyber Threats*. New York: Penguin Press, 2019.

Daniel, Michael. "Heartbleed: Understanding when we disclose cyber vulnerabilities." *White House Blog*, 28 April 2014. https://obamawhitehouse.archives.gov/blog/2014/04/28/heartbleed-understanding-when-we-disclose-cyber-vulnerabilities.

Department of Defense. *Network Centric Warfare – Creating a Decisive Warfighting Advantage*. Washington, D.C.: Department of Defense, 2003.

Ellman, Jesse, Lisa Samp, and Gabriel Coll. *Assessing the Third Offset Strategy*. Washington, D.C.: Center for Strategic & International Studies, 2017.

Fischerkeller, Michael P., and Richard J. Harknett. "Deterrence is not a credible strategy for cyberspace." *Orbis* 61, no. 3 (2017): 381–393.

Gartzke, Erik, and Jon R. Lindsay. "Weaving tangled webs: Offense, defense, and deception in cyberspace." *Security Studies* 24, no. 2 (2015): 316–348.

Geist, Edward Moore. "It's already too late to stop the AI arms race—we must manage it instead." *Bulletin of the Atomic Scientists* 72, no. 5 (2016): 318–321. https://doi.org/10.1080/00963402.2016.1216672.

Harknett, Richard J., and Max Smeets. "Cyber campaigns and strategic outcomes." *Journal of Strategic Studies* (March 2020): 1–34.

Healey, Jason. "The US government and zero-day vulnerabilities: From pre-heartbleed to shadow brokers." *Columbia Journal of International Affairs* (November 2016).

Healey, Jason. "The implications of persistent (and permanent) engagement in cyberspace." *Journal of Cybersecurity* 5, no. 1 (2019): 1–15.

Horowitz, Michael C. "Artificial intelligence, international competition, and the balance of power." *Texas National Security Review* 1, no. 3 (2018): 37–57. https://doi.org/10.15781/T2639KP49.

Horowitz, Michael C., Gregory C. Allen, Edoardo Saravalle, Anthony Cho, Kara Frederick, and Paul Scharre. *Artificial Intelligence and International Security*. Washington, D.C.: Center for New American Security, 2018.

Jacobsen, Jeppe T. "Lacan in the US cyber defence: Between public discourse and transgressive practice." *Review of International Studies* 46, no. 5 (2020): 613–631.

Jahnke, Art. "Could DeepCode AI make life harder for hackers?" *Boston University – The Brink*, 11 December 2018. http://www.bu.edu/articles/2018/deepcode-artificial-intelligence/.

Jensen, Benjamin M., Christopher Whyte, and Scott Cuomo. "Algorithms at war: The promise, peril, and limits of artificial intelligence." *International Studies Review* 25 (June 2019): 1–25.

Johnson, James. "Artificial intelligence & future warfare: Implications for international security." *Defense & Security Analysis* 35, no. 2 (2019): 147–169.

Johnson, James. "Artificial Intelligence: A Threat to Strategic Stability." *Strategic Studies Quarterly* 14, no. 1 (2020): 16–39.

Joiner, Keith F., and Malcom G. Tutty. "A tale of two allied defence departments: New assurance initiatives for managing increasing system complexity, interconnectedness and vulnerability." *Australian Journal of Multi-Disciplinary Engineering* 14, no. 1 (2018): 4–25.

Joyce, Rob. "Disrupting nation state hackers." *USENIX Enigma*, 27 January 2016. https://www.usenix.org/node/194636.

Joyce, Rob. "Improving and making the vulnerability equities process transparent is the right thing to do." *Whitehouse.Gov* (blog), 15 November 2017. https://www.whitehouse.gov/articles/improving-making-vulnerability-equities-process-transparent-right-thing/.

Kania, Elsa B. *Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power*. Washington D.C.: Center for New American Security, 2017.

Kello, Lucas. *The Virtual Weapon and International Order*. New Haven and London: Yale University Press, 2017.

Klare, Michael T. "AI arms race gains speed." *Arms Control Today* 49, no. 2 (2019): 35.

Knake, Robert, and Ari Schwartz. "Government's role in vulnerability disclosure: Creating a permanent and accountable vulnerability equities process." Discussion paper, Cyber Security Project, Belfer Center, June 2016.

Krepinevich, Andrew F. *The Military-Technical Revolution: A Preliminary Assessment*. Washington, D.C.: Center for Strategic and Budgetary Assessments, 2002.

Lindsay, Jon R. "Stuxnet and the limits of cyber warfare." *Security Studies* 22, no. 3 (2013): 365–404.

Maddison, John. "Using advanced AI to stay ahead of cybercriminals." *Infradata* (blog), 29 August 2019. https://www.infradata.com/news-blog/using-advanced-ai-to-stay-ahead-of-cybercriminals/.

Martinage, Robert. *Towards a New Offset Strategy – Exploiting U.S. Long-Term Advantages to Restore U.S. Global Power Projection Capability*. Washington, D.C.: Center for Strategic and Budgetary Assessments, 2014.

Mazanec, Brian M. *The Evolution of Cyber War: International Norm for Emerging-Technology Weapons*. Potomac Books/University of Nebraska Press, 2015.

McCardle, Jennifer. *Victory Over and Across Domains: Training for Tomorrow's Battlefield*. Washington, D.C.: Center for Strategic and Budgetary Assessments, 2019.

Mori, Satoru. "US defense innovation and artificial intelligence." *Asia-Pacific Review* 25, no. 2 (2018): 16–44.

Nakashima, Ellen, and Joby Warrick. "For NSA chief, terrorist threat drives passion to "collect it all." *Washington Post*, 14 July 2013. https://www.washingtonpost.com/world/national-security/for-nsa-chief-terrorist-threat-drives-passion-to-collect-it-all/2013/07/14/3d26ef80-ea49-11e2-a301-ea5a8116d211_story.html?utm_term=.d6d4eff5d534.

Nakashima, Ellen, and Reed Albergotti. "The FBI wanted to unlock the San Bernardino shooter's iPhone. It turned to a little-known Australian firm." *Washington Post*, 14 April 2021. https://www.washingtonpost.com/technology/2021/04/14/azimuth-san-bernardino-apple-iphone-fbi/.

Nakasone, Paul M., and Michael Sulmeyer. "How to compete in cyberspace." *Foreign Affairs*, 25 August 2020. https://www.foreignaffairs.com/articles/united-states/2020-08-25/cybersecurity.

National Academy of Science. *Computers at Risk: Safe Computing in the Information Age*. Washington, D.C.: The National Academies Press, 1991. https://doi.org/10.17226/1581.

National Security Commission on Artificial Intelligence. *Final Report*. 2021. https://assets.foleon.com/eu-west-2/uploads-7e3kk3/48187/nscai_full_report_digital.04d6b124173c.pdf.

Nye, Joseph S. *Cyber Power*. Cambridge, MA: Belfer Center for Science and International Affairs, Harvard University, 2010.

Owens, William A., and Edward Offley. *Lifting the Fog of War*. Baltimore, MD: Johns Hopkins University Press, 2001.

Pellerin, Cheryl. "Deputy secretary: Third offset strategy bolsters America's military deterrence." *U.S. Department of Defence*, 31 October 2016. https://www.defense.gov/Explore/News/Article/Article/991434/deputy-secretary-third-offset-strategy-bolsters-americas-military-deterrence/.

Perlroth, Nicole. *This Is How They Tell Me the World Ends: The Cyberweapons Arms Race*. New York: Bloomsbury Publishing, 2021.

Rid, Thomas. *Cyber War Will Not Take Place*. Oxford: Oxford University Press, 2013.

Roff, Heather M. "The frame problem: The AI "arms race" isn't one." *Bulletin of the Atomic Scientists* 75, no. 3 (2019): 95–98. https://doi.org/10.1080/00963402.2019.1604836.

Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War*. New York: W.W. Norton & Company, 2018.

Scharre, Paul. "Killer apps: The real dangers of an AI arms race." *Foreign Affairs* 98, no. 2 (2019): 135–144.

Schneider, Jacquelyn. *Digitally-Enabled Warfare: The Capability-Vulnerability Paradox*. Washington, D.C.: Center for New American Security, 2016.

Schneider, Jacquelyn. "The capability/vulnerability paradox and military revolutions: Implications for computing, cyber, and the onset of war." *Journal of Strategic Studies* 42, no. 6 (2019): 841–863.

Schneier, Bruce. *The Coming of AI Hackers*. Cambridge, MA: Belfer Center for Science and International Affairs, 2021.

Seymour, John, and Philip Tully. "Weaponizing Data Science for Social Engineering: Automated E2E Spear Phishing on Twitter." Conference paper, Black Hat USA 2016, 4–7 August 2016, Las Vegas, NV, USA. https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf.

Sheridan, Kelly. "How attackers use machine learning to predict BEC success." *DARKReading*, 26 July 2017. https://www.darkreading.com/vulnerabilities---threats/how-attackers-use-machine-learning-to-predict-bec-success/d/d-id/1329475.

Simonite, Tom. "For superpowers, artificial intelligence fuels new global arms race." *Wired*, 8 September 2017. https://www.wired.com/story/for-superpowers-artificial-intelligence-fuels-new-global-arms-race/.

Slayton, Rebecca. "What is the cyber offense-defense balance? Conceptions, causes, and assessment." *International Security* 41, no. 3 (2017): 72–109.

Stevens, Tim. "Cyberweapons: Power and the governance of the invisible." *International Politics* 55 (2018): 482–502.

Stoecklin, M.P., J. Jang, and D. Kirat. "Deeplocker: How AI can power a stealthy new breed of malware." *Security Intelligence*, 8 August 2018. https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/.

Sulmeyer, Michael, and Kathryn Dura. "Beyond killer robots: How artificial intelligence can improve resilience in cyber space." *War on the Rocks* (blog), 6 September 2018. https://warontherocks.com/2018/09/beyond-killer-robots-how-artificial-intelligence-can-improve-resilience-in-cyber-space/.

The White House, Office of Science and Technology Policy. "Artificial intelligence initiative." *Year One Annual Report,* 2020. https://www.whitehouse.gov/wp-content/uploads/2020/02/American-AI-Initiative-One-Year-Annual-Report.pdf.

Toffler, Alvin, and Heidi Toffler. *War and Anti-War: Survival at the Dawn of the 21st Century*. London: Warner Books, 1994.

Trump, Donald J. "National cyber strategy of the United States of America." *The White House*, 2018. https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/09/National-Cyber-Strategy.pdf.

Trump, Donald J. "Executive order on maintaining American leadership in artificial intelligence." *The White House*, 2019. https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/.

U.S. DoD. "Summary of the 2018 department of defense artificial intelligence strategy – harnessing AI to advance our security and prosperity." *U.S. Department of Defense*, 2018. https://man.fas.org/eprint/dod-ai.pdf.

US State Senate. *Hearing to Receive Testimony on U.S. Strategic Command, U.S. Transportation Command, and U.S. Cyber Command in Review of the Defense Authorization Request for Fiscal Year 2016 and the Future Years Defense Program*. Washington D.C.: US State Senate, 2015. https://www.armed-services.senate.gov/hearings/15-03-19-us-strategic-command-us-transportation-command-and-us-cyber-command.

USCYBERCOM. "U.S. cyber command factsheet." 16 April 2014. https://web.archive.org/web/20140416192156/http://www.stratcom.mil/factsheets/2/Cyber_Command/.

USCYBERCOM. "Achieve and maintain cyberspace superiority." *Command Vision for US Cyber Command*. 2018. https://www.cybercom.mil/Portals/56/Documents/USCYBERCOM%20Vision%20April%202018.pdf?ver=2018-06-14-152556-010.

Whyte, Christopher. "Poison, persistence, and cascade effects." *Strategic Studies Quarterly* 14, no. 4 (Winter 2020): 18–46.

Whyte, Christopher, and Brian Mazanec. *Understanding Cyber Warfare: Politics, Policy and Strategy*. Routledge, 2019.

Wilner, Alex S. "US cyber deterrence: Practice guiding theory." *Journal of Strategic Studies* 43, no. 2 (2020): 245–280.

Work, Robert O. "Remarks by deputy secretary work on third offset strategy." *Speech, Brussels, Belgium*, 28 April 2016. https://www.defense.gov/Newsroom/Speeches/Speech/Article/753482/remarks-by-deputy-secretary-work-on-third-offset-strategy/.

Yu, Ning, Zachary Tuttle, Carl Jake Thurnau, and Emmanuel Mireku. "AI-powered GUI attack and its defensive methods." *Proceedings of the 2020 ACM Southeast Conference*, Session 1 – Full papers, 2–4 April 2020. https://dl.acm.org/doi/pdf/10.1145/3374135.3385270.

Part III

Normative and legal challenges

7 Ethical principles for artificial intelligence in the defence domain[1]

Mariarosaria Taddeo, David McNeish, Alexander Blanchard and Elizabeth Edgar

Introduction

In information societies, maintaining a technological advantage is pivotal to the success of national defence and security measures. This is why over the past two decades there have been growing efforts to design, develop, and deploy digital technologies in this domain. Efforts span from the internet of things (IoT) to robotics and artificial intelligence (AI).

AI, in particular, has shown to have great potential to aid national defence and security practices. This technology has become efficient and effective in addressing complex and important tasks; both within the civil and military domain. Indeed, scholars, policy-makers and military experts observe that there is an on-going global race for the development of AI as a defence and security capability. For example, the latest national defence and innovation strategies of several governments – UK,[2] US,[3] Chinese,[4] Singapore,[5] Japanese,[6] and Australian[7] – explicitly mention AI capabilities, which are already deployed to improve the security of critical national infrastructures, such as transport, hospitals, energy and water supply.

The possible applications of AI in national defence and security are virtually unlimited, ranging from support to logistics and transportation systems to target recognition, combat simulation, training, and threat monitoring. This potential is coupled with serious ethical challenges. If left unaddressed, these challenges could hinder the adoption of AI for national defence and security or pose significant problems for our societies, like escalation of conflicts, the promotion of mass surveillance measures, as well as the spreading of misinformation or breaches of individual rights.[8] However, these ethical challenges are serious and hard to overcome but they can be addressed successfully,[9] if the design, development, and use of AI are informed by ethical considerations and guidance.

The goal of this document is to offer guidance by identifying ethical principles to inform the design, development, and use of AI for defence and security purposes.

These principles should not be taken as an alternative to national and international laws; rather they offer guidance to the use of AI in the defence

DOI: 10.4324/9781003284093-10

and security domain in ways that are coherent with existing regulations. In this sense, these principles indicate what ought to be done or not to be done "*over and above* the existing regulation, not against it, or despite its scope, or to change it, or to by-pass it (e.g. in terms of self-regulation)."[10]

As we shall see in the section "Ethical Guidelines for the Use of AI", ethical guidelines for the use of AI are often designed to be coherent with values of the organisation or key constitutional values of the country issuing the guidelines. For example, the principles defined by the US Defence Innovation Board (DIB) rest on International Humanitarian Law, as well as on core values of the US Armed Forces.[11] The European Group on Ethics in Science and New Technologies points towards EU Treaties and the EU Charter of Fundamental Rights as a starting point for the development of ethical values.[12]

In the rest of this report, the second section describes the methodology used for this analysis. The third section offers a definition of AI and an analysis of the ethical problems linked to current uses of this technology for defence and security. The fourth section introduces five principles to guide the specific deployment of AI for national defence and security. The fifth section concludes the chapter.

Methodology

The first step to identifying viable ethical principles to guide the use of AI for national defence and security is the identification of the ethical problems that this use may pose and that the principles should address. However, the choice as to how to identify these problems is not a trivial one. One may think of developing a complete taxonomy of ethical issues of AI in defence and security; but this is unfeasible and of little value: the taxonomy would be quickly outdated by the rapid developments in AI and its application to new uses. At the same time, different ethical problems may become evident when considering AI from different points of view. For example, some ethical problems of AI are inherent to the design and development processes, others emerge with the specific domain and purpose of deployment. Hence, the choice of level of analysis (or levels of abstraction) becomes crucial. Analyses that disregard the specific domain and purpose of deployment risk defining ethical principles which are too generic to provide any concrete guidance. At the same time, analyses that try to address all possible ethical challenges related to the use of AI in a specific domain risk losing sight of the need to harmonise ethical principles for uses of AI in one domain with the broader set of values underpinning our societies.

To avoid both these risks, our analysis relied on the method of the *Levels of Abstraction* (LoAs)[13] to identify the ethical problems related to the use of AI in defence and security. Before delving into our analysis, let us introduce LoAs.

LoAs are used in Systems Engineering and Computer Science to design models of a given system.[14] They are also widely used in Digital Ethics and have

been applied in this field of research to address several key issues, like identifying the responsibilities of online service providers,[15] offering guidance on the deployment of tracing and tracking technologies during the COVID-19 pandemic,[16] analysing the possibilities of deterrence in cyberspace,[17] or considering the ethical implications of trust in digital technologies.[18]

The method starts from the assumption that any system can be observed by focusing on specific properties while disregarding others. The choice of these properties, i.e. the observables, depends on the observer's aim. For example, for an engineer interested in maximising the aerodynamics of a car, the observables may be the shape of its parts, their weight and the materials. For a customer interested in the aesthetics of the car, the observables may be instead its colour, the car's interiors, and the overall look. The engineer and the customer observe the same car (system) at different LoAs, which will enable them to define different models of the car.

Thus, a LoA is defined as a finite but non-empty set of observables accompanied by a statement of what feature of the system under consideration such a LoA stands for. It is important to stress that a LoA does not reduce a car to merely the aerodynamics of its parts or to its overall look. Rather, a LoA is a tool that helps to make explicit the system observation perspective and constrain it only to those elements that are functional in a particular observation for the chosen aim.[19] LoAs can be organised in a gradient of abstractions (GoA), this is a way to consider a range of LoAs. A LoA can have a lower or higher granularity.

> The quantity of information in a model varies with the LoA: a lower LoA, of greater resolution or finer granularity, produces a model that contains more information than a model produced at a higher, or more abstract, LoA.[20]

When considering the ethical challenges of AI used in national defence and security one may focus on different LoAs, for example one may decide to consider only ethical problems emerging during the design stage of this technology and disregard the development and deployment steps. Similarly, ethical analyses may focus only on the intention of use or only on the effects of use of AI in this domain. Given the goal of this document, we choose a GoA that combines two LoAs: $\text{LoA}_{\text{purpose}}$ and $\text{LoA}_{\text{ethics}}$. The observables of $\text{LoA}_{\text{purpose}}$ are the immediate purposes (henceforth: purpose) of deployment of AI. The observables of $\text{LoA}_{\text{ethics}}$ are, for any given purpose, the aspects of the design, development and deployment of AI that may lead to un/ethical consequences.

It is worth stressing that the purpose is not the *function* of a specific technological artefact, an artefact with the same function may pose different ethical problems when deployed for different purposes. For example, consider a hypothetical AI image recognition system working on different databases. In one implementation the system may be used to grant access to a facility, while in other it may be deployed to identify targets of the battlefield. While the

function – recognising images – of the system remains the same, the purpose changes and with it the ethical implications to consider. The choice to focus on purposes of use rather than on the function of the technology rests on two reasons: the malleability of digital technologies and the goal of the analysis that we provide in this chapter. *Malleability* refers to the fact that digital technologies, even the most sophisticated, can easily be repurposed. As Moor put it:

> [Digital technologies] can be shaped and molded to do any activity that can be characterized in terms of inputs, outputs, and connecting logical operations. Logical operations are the precisely defined steps which take a computer from one state to the next. The logic of [digital technologies] can be massaged and shaped in endless ways through changes in hardware and software.[21]

Because of their malleability, un/ethical implications of digital technologies, AI in particular, are not necessarily defined by their design function as much as they are determined by the purpose with which these technologies are deployed. Within the defence and security domain these purposes can be clearly identified and are likely to shape both current and future uses of AI, thus they can inform the identification of ethical implications of the current and future uses of AI in this domain.

The goal of this analysis is not to define a comprehensive taxonomy of AI technologies and their related ethical implications, but to offer criteria to identify the ethical challenges linked to use of AI in the defence and security domain and provide ethical guidance to address them, this is why the focus on the purpose of the use is key. In this sense, the reader should consider the rest of this document not so much as a map of the possible uses and related ethical implications, but as a compass to orient practitioners and researchers working on the ethics of AI for national defence and security.

The LoAs embraced for this analysis have a medium granularity, as they focus on specific purpose of deployment in the domain of defence and security. Thus, they identify problems (and inform the definition of principles) that are not directly applicable to other domains, e.g. healthcare or public policy. At the same time, the LoAs abstract from specific contexts (e.g. naval or aviation) of AI deployment within the defence and security domain and disregard the variation of ethical challenges that may occur between different contexts. Consider, for example, the different problems and related solutions for using AI to aid submarines or aviation operations.

Using this GoA we identified three purposes of deployment of AI in defence and security: sustainment and support, adversarial and non-kinetic, adversarial and kinetic. We shall delve into these categories in the next section, but let us describe them briefly here. Sustainment and support uses of AI refer to all cases in which AI is deployed with the purpose to support 'back-office' functions, as well as logistics distribution of resources. This category also includes uses of AI to improve the security of infrastructures and

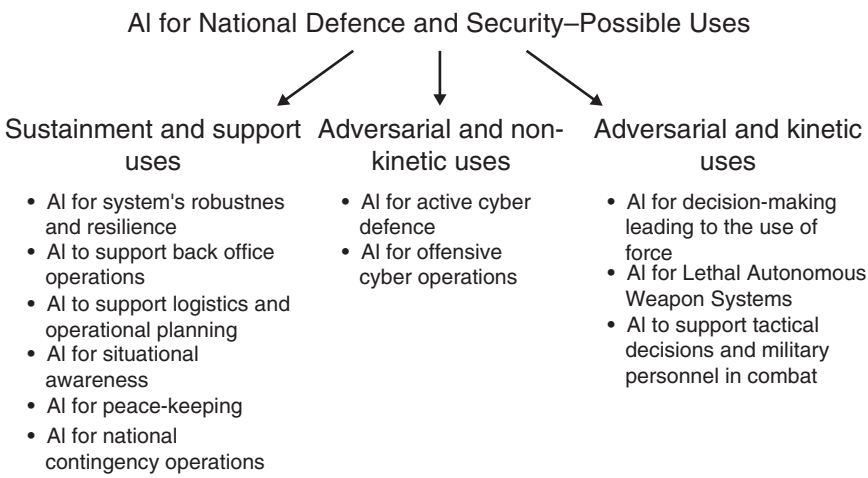

Figure 7.1 The three purposes of use of AI for national defence and security.
Source: Author's creation.

communication systems underpinning national defence and security services. Adversarial and non-kinetic uses of AI range from uses of AI to counter cyber-attacks to active cyber defence, and aggressive cyber operations with non-kinetic aims. Adversarial and kinetic uses refer to the integration of AI systems in combat operations, these range from the use of AI systems for threat identification to lethal autonomous weapons systems (LAWS).

The ethical principles for the use of AI in the defence and security domain that we provide in the fifth section refer to the purposes of use defined in this section. It should be noted that, the principles defined in this chapter focus only on sustainment and support uses and on adversarial and non-kinetic uses. Adversarial and kinetic uses will be the focus of further work, which will build on the findings of this study.

Ethical challenges of AI for defence and security purposes

AI draws upon a variety of approaches, methods, and models for use across a broad range of purposes. For the goal of this chapter, however, we can abstract from specific technical aspects (we can disregard, for example, whether the system under analysis is a statistical or a subsymbolic one) to focus only on the features of AI systems from where ethical challenges arise. At this LoA, AI is described as

> a growing resource of interactive, autonomous, and self-learning agency, which can be used to perform tasks that would otherwise require human

> intelligence to be executed successfully. This definition identifies in AI applications a growing resource of interactive, autonomous, and self-learning *agency*, to deal with tasks that would otherwise require human intelligence and intervention to be performed successfully.[22]

This combination of autonomy and learning skills underpins both beneficial and ethically problematic uses of AI. When considering the latter, at a high LoA, five key challenges have been identified in the relevant literature.[23] These are:

- enabling human wrongdoing;
- reducing human control;
- removing human responsibility;
- devaluing human skills; and
- eroding human self-determination.

All five challenges are relevant for the use of AI in defence and security, but some – enabling human wrongdoing, reducing human control, and removing human responsibilities – are key to the case in point.

- 'Enabling human wrongdoing' refers to AI systems that foster undue biases in the decision-making processes that may lead to erroneous or unfair decisions.
- 'Reducing human control' is a pressing challenge given the lack of predictability of the outcomes of AI systems. This challenge becomes even more pressing when considering the lack of transparency (or opaqueness) and explainability of the processes of these systems. Opaqueness and lack of explainability hinder human control insofar as they make it hard to scrutinise how a given output has been produced, identifying and correcting errors, as well as auditing AI systems for unethical and unwanted consequences.
- 'Removing human responsibility' is problematic as transparency issues and distributed design and development processes make it difficult to identify the source of errors and unintended consequences of AI systems. In turn, this leads to a responsibility gap.
- Two more challenges have been identified in the relevant literature[24] with respect to the deployment of AI for defence and security purposes. These are: 'escalation' of activities and 'lack of control' (the red circles in Figure 7.2).
- **Escalation**. The risk of escalation emerges especially in context where AI is deployed to support cyber operations. In these cases, AI can refine strategies and launch more aggressive counter operations. This may snowball into an intensification of attacks and responses, which, in turn, may threaten key infrastructures of our societies.[25]

- **Lack of control**. It is worth stressing that the risk related to the lack of control should not be confused with any sci-fi scenarios, where machines escape human control and overtake humans. These unrealistic scenarios divert attention from more concrete and pressing problems, like maintaining meaningful control of AI systems, of the cascade effects that their use may have, and ascribing responsibilities when deploying pervasive, distributed systems, with multiple interactions, and opaque, fast-pace execution.

At a high LoA, ethical problems and related solutions (*desiderata)* of AI in defence and security may be mapped against these challenges. However, when considering the ethical challenges of AI at the $LoA_{purpose}$ and LoA_{ethics}, it becomes clear that the solutions to address them require a more granular approach to be effective. Figure 7.2 shows the ethical desiderata for each of purpose of use of AI in the defence and security domain defined in the previous section.

The three purposes of use of AI in the defence and security domain are more ethically problematic as one moves from sustainment and support uses to adversarial and kinetic uses. This is because alongside the ethical problems related to the use of AI (e.g. transparency and fairness) one also needs to consider the ethical problems related to adversarial, whether non-kinetic or kinetic, uses of this technology and its disruptive and destructive impact.

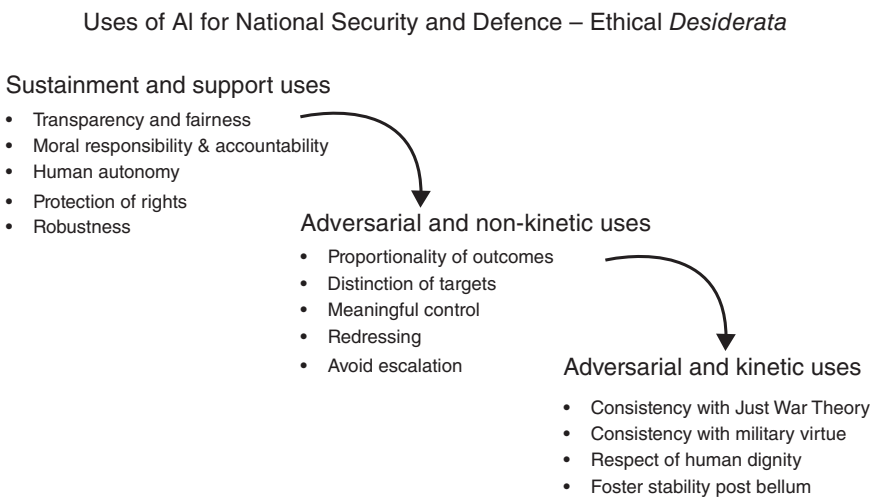

Figure 7.2 A map of the ethical desiderata linked to the specific purpose of the use of AI.

Source: Author's creation.

As shown in Figure 7.2, each category of use has its own specific ethical desiderata, but also inherits the ones from the categories on its left. For example, adversarial and non-kinetic uses of AI need to ensure the protection of rights, oversight and redressing while limiting the risks of escalation. To be ethically sound, these uses of AI need also to respect transparency and autonomy, which appear in the sustainment and support category. Similarly, adversarial and kinetic uses will have to ensure transparency and autonomy, alongside the protection of rights and proportionality, while also respecting the principles of Just War Theory, military virtue, human dignity and foster stability. Let us now consider in more details some of the key ethical challenges of each purpose of use.

Sustainment and support uses of AI

Defence organisations already employ AI systems for different non-aggressive aspects of operations.[26] Uses vary from applications in cybersecurity, where AI plays an ever-growing role to ensure systems robustness and resilience, to AI-based drones capturing video reconnaissance, radio-frequency identification (RFID) tags on food supply.[27]

For nations with adequate capabilities, AI systems are likely to reach full integration into national defence and security capabilities to support back-office, logistics and security tasks. For example, research estimates that the number of intelligent sensors in a military setting could reach one million per square kilometre similar to the supported connection density of the 5G network.[28] This has been described as *the internet of battle things*.[29] In these cases, AI will be used to ensure the robustness and resilience of the networks as well as to elaborate data and extract relevant information (*epistemic tasks*). All these uses pose serious ethical risks.

First, consider the use of AI to enhance system robustness. This refers to AI for software testing, which is a new area of research and development. It is defined as an "emerging field aimed at the development of AI systems to test software, methods to test AI systems, and ultimately designing software that is capable of self-testing and self-healing."[30]

AI can help with verification and validation of software, liberating human experts from tedious jobs, and offering a faster and more accurate testing of a given system.[31] In this sense, AI can take software testing to a new level, making systems more robust. However, we should be careful as societies about the way we use AI in this context, for delegating testing to AI could lead to a complete deskilling of defence personnel deployed for verification and validation of systems and networks and subsequent lack of control of this technology.

Next, let us focus on system resilience. AI is increasingly deployed for threat and anomaly detection (TAD). TAD can make use of existing security data to train for pattern recognition. For example, in April 2017 software firm DarkTrace launched Antigena, which uses machine learning to spot

abnormal behaviour on an IT network, shut off communications to that part of the system, and issue an alert. These services analyse malware and viruses, and some are able to quarantine threats and portions of the system for further investigation.

In certain cases, threat scanners have access to files, emails, mobile and endpoint devices, or even traffic data on a network. Monitoring extends to users as well. AI can be used to authenticate users by monitoring behaviour and generating biometric profiles, like for example, the unique way in which a user moves her mouse around.[32] Sometimes, this may imply tracking "sensor data and human-device interaction from your app/website. Every touch event, device motion, or mouse gesture is collected."[33] The risk is clear here. AI can improve system resilience to attacks but this requires extensive monitoring of the system and comprehensive data collection to train the AI. This poses users' privacy under a sharp devaluative pressure, exposing users to extra risks should data confidentiality be breached, and creating a mass-surveillance effect.[34] The sustainment and support uses of AI in defence and security pose ethical challenges similar to those related to uses of in AI in other domains, like for example the risks to breach privacy. This does not make these challenges less important. Indeed, a state actor breaching privacy poses severe risks to human right and fundamental democratic values. These problems need to be addressed with respect to their merit in the defence and security domain, which may complicate their solutions, for they involve balancing state interest, national security and respect of individual rights.

AI is also deployed to enhance situational awareness. Timely, situational awareness is decisive to enhance preparedness and to pre-empt threats. However, raising such an awareness using AI can be ethically challenging, especially given the hybrid nature of threats and different variables at play. This is because the threats, which may be hybrid in nature, may also coincide with changing facets in the political, economic, strategic, cultural, and social circumstances operating around the defender, and attacks can be initiated by actors working with changing allies, interests, resources and methods. This requires 'always-on,' real-time analytics and anomaly detection capabilities. AI offers much to this end, as it enables the analysis of great volumes of data. The key challenge is to ensure that large-scale data collection and analyses are kept in balance with key regulations and ethical values, to avoid undermining civilian trust in defence and security institutions, for example through excessive, undue surveillance or discriminatory systems.

AI can extract information to support logistics and decision-making, but also for foresight analyses, internal governance and policy. These are perhaps some of the uses of AI with the greater potential to improve defence and security operations, as they will facilitate timely and effective management of both human and physical resources, improve risk assessment, and support decision-making processes. For example, a recent chapter by KPMG stresses that a defence agency could have only a few minutes to decide whether a missile launch represents a threat, share the findings with allies, and decide

how to respond. AI would be of great help in this scenario, for it could integrate real-time data from satellites and sensors and elaborate key information that may contribute to the decision-making process. However, the challenge is that these uses of AI must ensure that AI systems would not perpetrate a biased decision and unduly discriminate, whilst also offering a means to maintain accountability and control and foster transparency.

Adversarial and non-kinetic uses of AI

The 2019 Global Risks Report of the World Economic Forum ranks cyber-attacks among the top five most likely sources of severe, global-scale risk.[35] The chapter is in line with other analyses about the escalation in frequency and impact of cyber-attacks,[36] and a Microsoft study shows that 60% of the attacks in 2018 lasted less than an hour and relied on new forms of malware.[37]

As the threats escalate, so does the need for defence strategies required to meet them. The UK and the US have employed 'active' cyber defence strategies that enable computer experts to neutralise or distract viruses with decoy targets, and to break back into a hacker's system to delete data or to destroy it completely. In 2016, the UK announced a £1.9 bn investment and a five-year plan to combat cyber threats. In February 2020, the UK also established the National Cyber Force, as a joint initiative between the Ministry of Defence and GCHQ, which is tasked to target hostile foreign actors. On an international scale, NATO can now rely on sovereign cyber effects in response to cyber-attacks, as agreed at the Brussels Summit.[38] This may enable the alliance to punish (attributed) attacks and deter attackers from striking again in the future.

AI will revolutionise these activities. Attacks and responses will become faster, more precise, and more disruptive. AI will expand the targeting ability of attackers, enabling them to use more complex and richer data. Enhancing current methods of attack is an obvious extension of existing technology; however, using AI within malware can change the nature and delivery of an attack. Autonomous and semi-autonomous cybersecurity systems endowed with a "playbook" of pre-determined responses to an activity, constraining the agent to known actions are already available on the market.[39] Autonomous systems able to learn adversarial behaviour and generate decoys and honeypots, thus actively luring threat actors,[40] are also being commercialised. Additionally, AI-enabled cyber weapons have already been prototyped including autonomous malware, corrupting medical imagery, and attacking autonomous vehicles.[41] For example, IBM created a prototype autonomous malware, DeepLocker, that uses a neural network to select its targets and disguise itself until it reaches its destination.[42]

As states use increasingly aggressive AI-driven strategies, opponents will respond ever more fiercely.[43] This may expand into an intensification of cyber-attacks and responses, which, in turn, may pose serious risks of

escalation and lead to kinetic consequences.[44] To avoid the escalation, it is vital that uses of AI respect key principles of Just War Theory which underpins international regulations, such as the United Nations Charter, [45] The Hague and Geneva Conventions,[46] and International Humanitarian Law,[47] and sets the parameters for both ethical and political debates on waging conflicts. It will be crucial that the deployment of AI for aggressive and non-kinetic purposes respects the principles of proportionality of responses, discriminates between legitimate and illegitimate targets, ensures some form of redressing when mistakes are made,[48] and maintains responsibility and control within the chain of command. Ultimately, ethical considerations on the adversarial and non-kinetic use of AI should contribute to understand how to apply Just War Theory in cyberspace and used to shape the debate on the regulation of in cyberspace.[49]

Adversarial and kinetic uses of AI

When considering the ethical implications of aggressive and kinetic uses of AI for defence and security, the main focus of analysis has been the combination of AI with machinery that can cause lethal harm to humans and destruction to physical objects in a completely autonomous way.

However, the use of AI for aggressive and kinetic purposes varies, ranging from automating various functions of a weapon system to systems that follow the pre-programmed instructions of a human, to full autonomy, when the weapons system will identify, select, and engage targets without any human input. Consider for example, a system developed for the Royal Navy called STARTLE[50] which supports human decision making with situational awareness software that monitors and assesses potential threats using a combination of AI techniques. Similarly, the Advanced Targeting & Lethality Automated System (ATLAS)[51] developed for the US Army support humans in identifying threats and prioritising potential targets. Ethical problems vary with the degree of autonomy and the ways in which AI might be involved in weapons systems, these go from ensuring that AI used to support human decision-making for the application of force works correctly to the level of autonomy of AWS and control exerted over them.

The UK government does not possess fully autonomous LAWS and has stated that it has no intention to develop them.[52] Political actors and military practitioners from other countries have expressed similar commitments. Nonetheless, it is important to consider and address the ethical challenges posed by fully autonomous LAWS to establish boundaries for the development and use of weapons which incorporate AI but are not fully autonomous in their operation.

A key challenge is to ensure that adversarial and kinetic uses of AI will be able to respect the tenets of Just War Theory, for example necessity, proportionality, and discrimination. So, for example, AI systems must be able to distinguish between a member of Armed Forces and a civilian carrying a

weapon or recognising the generally-accepted signs of surrender that operate in armed conflict. This may be problematic, because AI, at least in its current state of development, is insufficiently able to analyse context, in some situations its capacity to recognise who is and who is not a legitimate target could be significantly worse than that of humans.[53]

The responsibility gap is another key ethical challenge. As mentioned in Section "Ethical Challenges of AI for Defence and Security Purposes", whilst a responsibility gap is problematic in all the three categories of use of AI, it is particularly worrying when considering the adversarial and kinetic case, given the high stakes involved.[54] This gap becomes an even more pressing issue when coupled with the respect of an opponent and of her dignity. Treating opponents with respect in warfare is an important way of maintaining warfare's morality,[55] the interpersonal relation with the opponent is considered to be key to this end. Insofar as the use of autonomous LAWS would sever this relation, the question emerges as to whether the use of these systems undermines the dignity of those whom they target (and possibly also those who use them) and lead to a form of morally problematic killing.[56]

Finally, questions arise with respect to the impact of LAWS on international stability. On the one side, LAWS may reduce the time span of the hostilities in which states may engage and thus contribute to foster stability. They could also be an effective deterrent against possible opponents. On the other side, LAWS may lead to unjust war and hamper international instability. This is because the use of LAWS may lower the barriers to warfare[57] possibly increasing the number of wars. For instance, it may be the case that the widespread use of LAWS would allow decision-makers to wage wars without the need to overcome the potential objections of military personnel.[58] In the same vein, asymmetric warfare that would result from one side using LAWS may lead to the weaker side resorting to insurgency and terrorist tactics more often.[59] Because terrorism is generally considered to be a form of unjust warfare (or, worse, an act of indiscriminate murder), deploying LAWS may lead to a greater incidence of immoral violence.

Ethical guidelines for the use of AI

As argued in the previous section, the use of the AI in this domain poses serious ethical challenges which, if left unaddressed, may lead to disastrous consequences for national defence and security and for international stability. In this section, we offer five ethical principles, which build on the foundational bioethical principles identified by Floridi and Cowls[60] but are also specifically designed to address the ethical challenges linked to the deployment of AI in the defence and security domain. The principles specified in this chapter refer to both sustainment and support uses and adversarial and non-kinetic uses of AI. They should be regarded as the first building block of a more comprehensive ethical framework addressing also the adversarial and

kinetic uses of AI, which will be the focus of the second, forthcoming, part of this project.

In order to be ethically sound, sustainment and support and adversarial and non-kinetic uses of AI for national defence and security purposes should respect the following ethical principles:

I Justified and overridable uses
II Just and transparent systems and processes
III Human moral responsibility
IV Meaningful human control
V Reliable AI systems

Justified and overridable uses

The (non) adoption of AI needs to be justified to ensure that AI solutions are not being underused, thus creating opportunity costs; or overused and misused, thus creating risks. Similarly, the decision to (or not to) resort to AI should always be overridable, should it become clear that it leads to unwanted consequences.

Even when designed and deployed according to ethical principles, AI remains an ethically challenging technology. Its use may lead to great advantages for national defence and security. Yet, AI is not a silver bullet. This is a lesson that should be learned from the ethical governance of AI for social good. As Floridi and colleagues stress:

> it is important to acknowledge at the outset that there are myriad circumstances in which AI will not be the most effective way to address a particular social problem. This could be due to the existence of alternative approaches that are more efficacious or because of the unacceptable risks that the deployment of AI would introduce.[61]

At the same time, AI can also encroach upon human rights, International Humanitarian Law or pose risks to international stability (the reader will recall the risks of snowball effect linked to the adversarial and non-kinetic use of AI). This is why the decision to (or not to) delegate tasks to AI systems should follow a careful analysis of the risks and benefits in any given context of deployment to justify it.

This principle yields different recommendations when considering sustainment and support and adversarial and non-kinetic uses. In the first case, the principle calls for an assessment of the ethical risks against the expected benefits following from the deployment of AI systems. For example, weighting the benefits of using an AI system that may speed-up a decision-making process or optimise logistic and distribution of resources against the likelihood that it may have a negative impact on jobs and human expertise; or

considering the impact on human autonomy when AI is integrated in human teams (human-machine teaming).

When deciding on deploying AI for adversarial and non-kinetic purposes, for example for offensive cyber operations, it is essential to ensure that AI systems will respect the principles of necessity, humanity, distinction, and proportionality.[62] This may prove to be a complex task, as the principles of International Humanitarian Law are geared towards kinetic forms of war waging and therefore their implementation to the case of non-kinetic warfare may be problematic. Consider for example, proportionality and the problems of assessing the expected damage to intangible entities (e.g. data or services) against the concrete military aim to be achieved.[63] Satisfying this principle will require extending the scope of the fundamental tenets of Just War Theory from kinetic to non-kinetic war waging. A complex but necessary, and not impossible, task.

Given the learning capability of AI and the lack of predictability of its outcome, even when uses of AI are justified, a constant monitoring of the ethical soundness of the solutions that they provide should be in place. Similarly, procedures to override the decision to resort to AI in a timely and effective way should be established every time an AI system is deployed.

Just and transparent systems and processes

AI systems should not perpetrate any undue discrimination, nor should they lead to any breach of the principles of Just War Theory. This is why defence and security institutions should ensure that the deployed AI systems, and the processes in which they are embedded, remain transparent (and explicable) to facilitate the identification of the origin of unintended and mistaken outcomes, the attribution of responsibilities, and guarantee the possibility of scrutinising and challenging processes and outcomes to ensure that they remain ethically sound.

Three aspects are vital to this end:

- establish processes for ethical auditing;
- ensure that procured AI systems respect ethical principles;
- maintain traceability for the design, development or procurement, and deployment of AI systems.

Ethical auditing should involve the entire decision-making process, and so it should focus on both human agents and the technological systems, to ensure that both agents respect the relevant ethical principles. Transparency of AI systems and processes enables access to the relevant information. The former requires explainability, while the latter traceability.

Transparency of AI follows from the effort of designing and developing explainable technologies. Thus, it is crucial that in-house and procured AI systems are designed and developed with explainability in mind. Defence

and security agencies should consider participating actively in the 'design-develop-deploy' cycle of the AI technologies that they procure and contribute to the development phase by setting standards and offering a trusted space where these technologies could be beta-tested. To facilitate this process, *procurement* policies should account for an ethical scrutiny of the third parties involved.

AI systems are often designed and developed in a distributed way, models, data, training and implementation may be managed by different actors. At the same time, AI learns by experience: past deployments impact future outcomes. This is why transparency requires traceability of sourcing and practices, to ensure that the chain of events leading to possible unwanted outcomes is not lost in the distributed and dynamic nature of design, development and deployment of AI.

Human moral responsibility

Humans remain the only agents morally responsible for the outcomes of AI systems deployed for defence and security purposes. While AI systems can be considered moral agents, insofar as they perform actions that have a moral value,[64] they cannot be held morally responsible for those actions. This is because they lack intentionality and understanding of the reward/punishment that may result as a consequence of actions. Unplugging an AI system because it violated the principle of proportionality, for example, is not a punishment for the system.

However, ascribing responsibilities to humans for the actions of AI systems has proved to be problematic, due to the distributed and interconnected ways in which AI is developed and the lack of transparency and predictability of its outcomes. Two approaches can be followed to enable fair processes to ascribe responsibilities:

1 following the chain of command, control and communication;
2 faultless, back-propagation approach.

They can be described more simply as a 'linear' and a 'radial' approach, respectively. These two approaches are complementary and serve the twin purposes of addressing unwanted consequences, mis- and overuses of AI and to foster a self-improving dynamic in the network of agents involved in the design, development, and deployment of AI for defence and security.

According to the linear approach, responsibility is attributed following the chain of command, control and communication. In this case, decision-makers are held responsible for the unwanted consequences of AI, whether these results from failures of AI systems, unpredictability of outcomes or bad decisions. In order to ascribe responsibility fairly, it is essential that the decision-makers have adequate information and *understanding* of the way the specific AI system works in the given context, of its robustness, of

the risks that it may deliver unpredicted (and unwanted) outcomes, of the required level of meaningful control, and of the dangers that may follow if the AI systems fails to behave according to expectations. The linear approach entails a certain epistemic threshold. This means that for the use of AI must be coupled with proper training of the personnel both those who decide to deploy and those who use it so that they understand the ways in which AI systems work, risks and benefits linked to the systems, and the ethical and legal implications of the decision to deploy AI. This approach rests on the idea that informed decision-makers choosing to use AI do so while being aware of the risks that this may imply and take responsibility for it. This awareness, in a military context, can help to fill the so-called 'responsibility gap' of AI.

The radial approach is useful to address unwanted outcomes of AI systems that do not stem from bad intentions or follow from actions that are morally neutral per se. This approach addresses unethical consequences that spur from the convergence of different, independent, morally neutral factors. In the relevant literature this has been defined as *faultless responsibility*.[65] It refers to contexts in which, while it is possible to identify the causal chain of agents and actions that led to a morally good/bad outcome, it is not possible to attribute intent to perform morally good/bad actions to any of those agents individually and, therefore, all the agents are held morally responsible for that outcome insofar as they are part of the network which determined it.

This is not an entirely new approach, as it is akin to the legal concept of strict liability. According to strict liability, legal responsibility for unwanted outcomes is attributed to one or more agents for the damage caused by their actions or omissions, irrespective of the intentionality of the action and feasibility of control. When considering human-machine teaming – the integration of AI systems in defence and security infrastructures, decision-making processes, and operations – what one needs to show to attribute moral responsibility according to the radial approach is that

> some evil has occurred in the system, and that the actions in question caused such evil, but it is not necessary to show exactly whether the agents/sources of such actions were careless, or whether they did not intend to cause them.[66]

All the agents of the network are then held maximally responsible for the outcome of the network. The radial approach does not aim at distributing reward and punishment for the actions of a system, rather it aims at establishing a feedback mechanism that incentivises all the agents in the network to improve its outcomes – if all the agents are morally responsible, they may become more cautious and careful and this may reduce the risk of unwanted outcomes. This becomes quite effective when, for example, the moral responsibility is linked to the reputation of the agents.

Meaningful human control

It follows from the previous principle that the deployment of AI should also envisage meaningful forms of human control. These will be essential to limit the risks that the outcome of AI systems will not meet the original intent, to identify promptly mistakes and unintended consequences, as well as to ensure timely intervention on, or deactivation of, the systems, should this be necessary.

The concept of meaningful control has been discussed widely in the relevant literature on LAWS and indeed when considering these systems, control is a key element to consider. However, meaningful control is necessary also when considering uses of AI that may not lead to the use of force. This is because

> military systems must be able to function safely and effectively under a wide range of highly dynamic environments and use cases that are hard to predict or anticipate during the design phase. They must also be resilient to failure and to complex, uncertain and unpredictable events and situations where the dynamics of the military domain necessitate complex judgements regarding acceptable actions based on rules of engagement, international law and judgements over legality, proportionality and risk. Because of this the maintenance of Human Control through a combination of specification, design, training, operating procedures, and assurance processes is seen as critical in many, if not all military systems.[67]

Meaningful human control of AI is characterised as dynamic, multidimensional and situation dependent and it can be exercised focusing on different aspects of the human-machine team. For example, the Stockholm International Peace Research Institute and the International Committee of the Red Cross identify three main aspects of human control of weapon systems: the weapon system's parameters of use, the environment, and human-machine interaction.[68] More aspects can also be considered. For example, Boardman and Butcher suggest that control should not just be meaningful but 'appropriate,' insofar as it should be exercised in such a way to ensure that the human involvement in the decision-making process remains significant without impairing system performance.[69]

While meaningful control can be dynamic, multidimensional and situationally dependent; the principle that prescribes it is only effective insofar as it defines a lower threshold below which control is so minimal to become irrelevant. Hence, the principle can be implemented minimally and maximally. Minimally, the implementation of this principle requires having a human *on* the loop able to understand the functioning of the system and its implications and with the ability to 'unplug' the system timely and

effectively. Maximally, the principle requires individuals in charge of AI systems to combine technical, legal and ethical training to ensure that the decision *to let the system work* is informed by all relevant dimensions and not a mere vetting of the system.

Therefore, the principle does not admit *fire and forget* uses of AI, as it considers control as an element which can be modulated with respect to a rigorous risk assessment of unintended consequences, and related negative impact on national defence and international stability. Where even lower levels of meaningful control cannot be complemented with these assessments, the use of AI systems is ethically unwarranted. It should be noted that the principle is best implemented when protocols for the attribution of responsibilities for misuses of AI and mistakes made by AI systems are in place alongside effective redressing and remedy processes. The attribution of responsibility hinges on the respect of transparency.

Reliable AI systems

The principle mandates the establishing of meaningful monitoring of the execution of the tasks delegated to AI. The monitoring should be adequate to the learning nature of the systems, and their lack of transparency, while remaining feasible in terms of resources, especially time, and hence computational feasibility.

AI has a poor shock response (robustness) and any slight alterations to inputs can degrade a model disproportionately.[70] Thus, deploying on AI for defence and security purposes could favour opponents[71] if the system is not deployed according to procedure that envisage forms of monitoring and prompt redressing in case of mistakes. This is why this principle prescribes the deployment of reliable AI systems, that is systems which are being monitored throughout their deployment.

Forms of control may span from new forms of procurement that envisage an active role of the defence and security institutions in the design and development process; in house design and development of models; use of data for system training and testing collected, curated and validated by the systems providers directly and maintained securely; mandatory forms of adversarial training with appropriate levels of refinement of models to test their robustness; sparring training of AI models; monitoring the output of AI systems deployed in the wild with some form of *in silico* baseline model, as suggested by Taddeo, McCutcheon, and Floridi.[72]

As stressed in the methodology section of this chapter, AI systems are autonomous, self-learning agents interacting with the environment. Their behaviour depends as much on the inputs they are fed and interactions with other agents once deployed as it does on their design and training. Responsible uses of AI for defence and security purposes need to take into account the autonomous, dynamic, and self-learning nature of AI systems, and start

envisaging forms of monitoring that span from the design to the deployment stages.

Conclusion

These principles should not be followed as an algorithm, they do not offer a set of instructions that ensure ethically sound decisions with respect to the deployment of AI, rather they offer guidelines to spur and articulate ethical considerations with respect to the uses of AI in defence and security. For them to be effective, it is crucial that the principles are officially adopted by defence and security organisations and that a committee or supervising body in charge of fostering the adoption of these principles is established. At the same time, it is key to train members of staff with respect to the ethical implications of AI.

Different trade-offs among these principles will have to be defined depending on the context of deployment. Clearly the trade-offs will have to be coherent with the aim to minimise unethical consequences, with the ethical values imbued in the practices of defence and security institutions, with requirements set by laws and regulations, and ultimately with the values and rights underpinning our democracies. This is only possible (i.e. the trade-offs will be correct) insofar as the humans making the decision are able to take into account the principles offered in this document, along with knowledge of legal and technical aspects of AI with the goal to reconcile different values, interests, and goals.

Acknowledgement

We are very grateful to Isaac Taylor for his work and comments on an early version of this article and to Rebecca Hogg and the participants of the 2020 Dstl AI Fest for their questions and comments, for they enabled us to improve several aspects of our analysis. We are responsible for any remaining mistakes.

Funding information

Mariarosaria Taddeo and Alexander Blanchard's work on this article has been funded by the Dstl Ethics Fellowship held at the Alan Turing Institute. The research underpinning this work was funded by the UK Defence Chief Scientific Advisor's Science and Technology Portfolio, through the Dstl Autonomy Programme, grant number R-DST-TFS / D026. This paper is an overview of UK Ministry of Defence (MOD) sponsored research and is released for informational purposes only. The contents of this chapter should not be interpreted as representing the views of the UK MOD, nor should it be assumed that they reflect any current or future UK MOD policy. The information contained in this chapter cannot supersede any statutory or contractual requirements or liabilities and is offered without prejudice or commitment.

Notes

1 The analysis presented in this chapter has been published in M. Taddeo et al., "Ethical principles for artificial intelligence in national defence," *Philosophy & Technology* 34 (2021): 1707–1729.
2 https://www.gov.uk/government/publications/future-force-concept-jcn-117.
3 https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.
4 H. Roberts et al., "The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation," *AI & Society* (June 2020).
5 https://www.csa.gov.sg/~/media/csa/documents/publications/singaporecybersecuritystrategy.pdf.
6 https://www.nisc.go.jp/eng/pdf/cs-senryaku2018-en.pdf.
7 https://www.dst.defence.gov.au/strategy/defence-science-and-technology-strategy-2030.
8 M. Taddeo, "The struggle between liberties and authorities in the information age," *Science and Engineering Ethics* (September 2014); M. Taddeo, "Three ethical challenges of applications of artificial intelligence in cybersecurity," *Minds and Machines* 29, no. 2 (2019).
9 L. Floridi and M. Taddeo, "What is data ethics?" *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, no. 2083 (2016).
10 L. Floridi, "Soft ethics and the governance of the digital," *Philosophy & Technology* 31, no. 1 (2018): 4.
11 Defence Innovation Board [DIB], "AI principles: Recommendations on the ethical use of artificial intelligence by the department of defence," 2019.
12 European Commission, "Statement on artificial intelligence, robotics and "autonomous" systems," *European Group on Ethics in Science and New Technologies*, 2018.
13 L. Floridi, "The methods of levels of abstraction," *Minds and Machines* 18, no. 3 (2008).
14 C.A.R. Hoare, "Notes on data structuring," in *Structured Programming* (Academic Press Ltd, 1972); D. Heath, D. Allum, and L. Dunckley, *Introductory Logic and Formal Methods* (Henley-on-Thames: Alfred Waller, 1994); A. Diller, *Z: An Introduction to Formal Methods* (Wiley & Sons, 1994); J. Jacky, *The Way of Z: Practical Programming with Formal Methods* (Cambridge University Press, 1997); P. Boca, *Formal Methods: State of the Art and New Directions* (London: Springer, 2014).
15 M. Taddeo and L. Floridi, "The debate on the moral responsibilities of online service providers," *Science and Engineering Ethics* (November 2015).
16 J. Morley et al., "Ethical guidelines for COVID-19 tracing apps," *Nature* 582 (2020).
17 M. Taddeo, "The limits of deterrence theory in cyberspace," *Philosophy & Technology* (2017).
18 M. Taddeo, "Trusting digital technologies correctly," *Minds and Machines* 27, no. 4 (2017).
19 Floridi, "The methods of levels."
20 Ibid., 315.
21 J.H. Moor, "What is computer ethics?" *Metaphilosophy* 16, no. 4 (1985): 269.
22 L. Floridi and J. Cowls, "A unitified framework of five principles for AI in society," *Harvard Data Science Review* (June 2019).
23 G.Z. Yang et al., "The grand challenges of science robotics," *Science Robotics* 3, no. 14 (2018).
24 Ibid.

25 Taddeo, "The limits of deterrence."
26 US Army, "Robotic and autonomous systems strategy," 2017.
27 R.J. Lysaght, R. Harris, and W. Kelly, "Artificial intelligence for command and control" (ANALYTICS INC WILLOW GROVE PA, 1988); P. Fraga-Lamas et al., "A review on internet of things for defense and public safety," *Sensors (Basel, Switzerland)* 16, no. 10 (2016); J. Schubert et al., "Artificial intelligence for decision support in command and control systems," *23rd International Command and Control Research & Technology Symposium "Multi-Domain C2"* (2018).
28 A. Kott, A. Swami, and B.J. West, "The internet of battle things" ArXiv:1712.08980 [Cs] (December 2017); International Telecommunications Union, "minimum requirements related to technical performance for IMT-2020 radio interface(s)," 2017.
29 Kott, Swami and West, "The internet of battle things."
30 www.aitesting.org.
31 T.M. King et al., "AI for testing today and tomorrow: Industry perspectives," in *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)* (IEEE, 2019).
32 "BehavioSec: Continuous authentication through behavioral biometrics," *BehavioSec*, 2019.
33 http://www.unbotify.com.
34 M. Taddeo, "Cyber security and individual rights, striking the right balance," *Philosophy & Technology* 26, no. 4 (2013); Taddeo, "The struggle between liberties."
35 World Economic Forum, "The global risks report 2018," *World Economic Forum*, 2018.
36 M. Taddeo and L. Floridi, "Regulate artificial intelligence to avert cyber arms race," *Nature* 556, no. 7701 (2018).
37 Microsoft Defender ATP Research Team, "Protecting the protector: Hardening machine learning defenses against adversarial attacks," 2018.
38 https://www.nato.int/docu/review/articles/2019/02/12/natos-role-in-cyberspace/index.html.
39 "DarkLight offers first of its kind artificial intelligence to enhance cybersecurity defenses," *Business Wire*, 26 July 2017.
40 "Acalvio autonomous deception," *Acalvio*, 2019.
41 Y. Mirsky et al., "CT-GAN: Malicious tampering of 3d medical imagery using deep learning," *ResearchGate* (2019); J. Zhuge et al., "Collecting autonomous spreading malware using high-interaction honeypots," in *Information and Communications Security* (Springer, 2007).
42 "DeepLocker: How AI can power a stealthy new breed of malware," *Security Intelligence* (blog), 8 August 2018.
43 Taddeo and Floridi, "Regulate artificial intelligence."
44 Taddeo, "The limits of deterrence."
45 https://www.un.org/en/sections/un-charter/un-charter-full-text/.
46 https://www.loc.gov/rr/frd/Military_Law/pdf/ASubjScd-27-1_1975.pdf.
47 https://www.icrc.org/en/doc/resources/documents/misc/57jm93.htm.
48 M. Taddeo, "Information warfare: A philosophical perspective," *Philosophy and Technology* 25, no. 1 (2012); M. Taddeo, "An analysis for a just cyber warfare," in *Fourth International Conference of Cyber Conflict* (Tallinn: NATO CCD COE and IEEE, 2012); M. Taddeo, "Just information warfare," *Topoi* (April 2014).
49 M. Taddeo, "On the risks of relying on analogies to understand cyber conflicts," *Minds and Machines* 26, no. 4 (2016); M. Taddeo, "Cyber conflicts and political power in information societies," *Minds and Machines* 27, no. 2 (2017).
50 https://www.roke.co.uk/products/startle.

51 https://breakingdefense.com/2019/03/atlas-killer-robot-no-virtual-crewman-yes/.
52 United Nations Office at Geneva, "Human Machine Touchpoints: The United Kingdom's Perspective on Human Control Over Weapon Development and Targeting Cycles," 2018.
53 N. Sharkey, "Saying "No!" To lethal autonomous targeting," *Journal of Military Ethics* 9, no. 4 (2010);
N. Sharkey, "Killing made easy: From joysticks to politics," in *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press, 2012); N.E. Sharkey, "The evitability of autonomous robot warfare." *International Review of the Red Cross* 94, no. 886 (2012); G. Tamburrini, "On banning autonomous weapons systems: From deontological to wide consequential reasons," in *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge University Press, 2016).
54 R. Sparrow, "Killer robots," *Journal of Applied Philosophy* 24, no. 1 (2007).
55 T. Nagel, "War and massacre," *Philosophy and Public Affairs* 1 (1972).
56 P. Asaro, "On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making," *International Review of the Red Cross* 94 no. 886 (2012); B. Docherty, "Shaking the foundations: The human rights implications of killer robots," *Human Rights Watch*, 2014; A. Sharkey, "Autonomous weapons systems, killer robots and human dignity," *Ethics and Information Technology* 21, no. 2 (2019); A.M. Johnson and S. Axinn, "The morality of autonomous robots," *Journal of Military Ethics* 12, no. 2 (2013); R. Sparrow, "Robots and respect: Assessing the case against autonomous weapon systems," *Ethics & International Affairs* 30, no. 1 (2016); M.E. O'Connell, "Banning autonomous killing: The legal and ethical requirement that humans make near-time lethal decisions," in *The American Way of Bombing: How Legal and Ethical Norms Change* (Cornell University Press, 2014); M. Ekelhof, "Moving beyond semantics on autonomous weapons: Meaningful human control in operation," *Global Policy* 10, no. 3 (2019).
57 C. Enemark, "Drones over Pakistan: Secrecy, ethics, and counterinsurgency," *Asian Security* 7, no. 3 (2011); D. Brunstetter and M. Braun, "From Jus Ad Bellum to Jus Ad Vim: Recalibrating our understanding of the moral use of force," *Ethics & International Affairs* 27, no. 1 (2013).
58 J. McMahan, "Foreword," in *Who Should Die? The Ethics of Killing in War* (Oxford University Press, 2013).
59 Sharkey, "Saying "No!";" Sharkey, "Killing made easy."
60 L. Floridi and J. Cowls, "A unitified framework of five principles for AI in society," *Harvard Data Science Review* (June 2019).
61 L. Floridi et al., "How to design AI for social good: Seven essential factors," *Science and Engineering Ethics* 26, no. 3 (2020): 1773.
62 "The UK and International Humanitarian Law 2018," n.d.
63 Taddeo, "Information warfare;" Taddeo, "An analysis for a just cyber warfare;" Taddeo, "Just information warfare."
64 L. Floridi and J.W. Sanders, "On the morality of artificial agents," *Minds and Machines* 14, no. 3 (2004).
65 L. Floridi, "Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, no. 2083 (2016).
66 Ibid., 8.
67 M. Boardman and F. Butcher, "An exploration of maintaining human control in AI enabled systems and the challenges of achieving it," *STO-MP-IST-178*, 2019: 2.
68 V. Boulanin et al., "Limits on autonomy in weapon systems: identifying practical elements of human control," *Stockholm International Peace Research Institute and the International Committee of the Red Cross*, 2020.

69 Boardman and Butcher, "An exploration of maintaining human control."
70 M. Rigaki and A. Elragal, "Adversarial deep learning against intrusion detection classifiers," 2017.
71 M. Brundage et al., "The malicious use of artificial intelligence: forecasting, prevention, and mitigation" ArXiv:1802.07228 [Cs] (February 2018); M. Taddeo, T. McCutcheon, and L. Floridi, "Trusting artificial intelligence in cybersecurity is a double-edged sword," *Nature Machine Intelligence* 1, no. 12 (2019).
72 Taddeo, McCutcheon and Floridi, "Trusting artificial intelligence."

Bibliography

"Acalvio autonomous deception." *Acalvio*, 2019. https://www.acalvio.com.

Asaro, P. "On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making." *International Review of the Red Cross* 94 no. 886 (2012): 687–709. https://doi.org/10.1017/S1816383112000768.

"BehavioSec: Continuous authentication through behavioral biometrics." *BehavioSec*, 2019. https://www.behaviosec.com/.

Boardman, M., and F. Butcher. "An exploration of maintaining human control in AI enabled systems and the challenges of achieving it." *STO-MP-IST-178*, 2019. https://www.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-IST-178/MP-IST-178-07.pdf.

Boca, P. *Formal Methods: State of the Art and New Directions*. London: Springer, 2014.

Boulanin, V., M. Peldán Carlsson, N. Goussac, and D. Davidson. "Limits on autonomy in weapon systems: Identifying practical elements of human control." *Stockholm International Peace Research Institute and the International Committee of the Red Cross*, 2020. https://www.sipri.org/publications/2020/other-publications/limits-autonomy-weapon-systems-identifying-practical-elements-human-control-0.

Brundage, M., S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe. "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation." ArXiv:1802.07228 [Cs] (February 2018). http://arxiv.org/abs/1802.07228.

Brunstetter, D., and M. Braun. "From Jus Ad Bellum to Jus Ad Vim: Recalibrating our understanding of the moral use of force." *Ethics & International Affairs* 27, no. 1 (2013): 87–106. https://doi.org/10.1017/S0892679412000792.

"DarkLight Offers first of its kind artificial intelligence to enhance cybersecurity defenses." *Business Wire*, 26 July 2017. https://www.businesswire.com/news/home/20170726005117/en/DarkLight-Offers-Kind-Artificial-Intelligence-Enhance-Cybersecurity.

"DeepLocker: How AI can power a stealthy new breed of malware." *Security Intelligence* (blog), 8 August 2018. https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/.

Defence Innovation Board [DIB]. "AI principles: Recommendations on the ethical use of artificial intelligence by the department of defence." 2019. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF.

Diller, A. *Z: An Introduction to Formal Methods*. 2nd ed., Chichester: Wiley & Sons, 1994.

Docherty, B. "Shaking the foundations: The human rights implications of killer Robots." *Human Rights Watch*, 2014. https://www.hrw.org/report/2014/05/12/shaking-foundations/human-rights-implications-killer-robots.

Ekelhof, M. "Moving beyond semantics on autonomous weapons: Meaningful human control in operation." *Global Policy* 10, no. 3 (2019): 343–348. https://doi.org/10.1111/1758-5899.12665.

Enemark, C. "Drones over Pakistan: Secrecy, ethics, and counterinsurgency." *Asian Security* 7, no. 3 (2011): 218–237. https://doi.org/10.1080/14799855.2011.615082.

European Commission. "Statement on artificial intelligence, robotics and "autonomous" systems." *European Group on Ethics in Science and New Technologies*, 2018. https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1.

Floridi, L. "The methods of levels of abstraction." *Minds and Machines* 18, no. 3 (2008): 303–329. https://doi.org/10.1007/s11023-008-9113-7.

Floridi, L. "Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, no. 2083 (2016): 20160112. https://doi.org/10.1098/rsta.2016.0112.

Floridi, L. "Soft ethics and the governance of the digital." *Philosophy & Technology* 31, no. 1 (2018): 1–8. https://doi.org/10.1007/s13347-018-0303-9.

Floridi, L, and J. Cowls. "A unitified framework of five principles for AI in society." *Harvard Data Science Review* (June 2019). https://doi.org/10.1162/99608f92.8cd550d1.

Floridi, L., J. Cowls, T.C. King, and M. Taddeo. "How to design ai for social good: Seven essential factors." *Science and Engineering Ethics* 26, no. 3 (2020): 1771–1796. https://doi.org/10.1007/s11948-020-00213-5.

Floridi, L., and J.W. Sanders. "On the morality of artificial agents." *Minds and Machines* 14, no. 3 (2004): 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Floridi, L., and M. Taddeo. "What is data ethics?" *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, no. 2083 (2016): 20160360. https://doi.org/10.1098/rsta.2016.0360.

Fraga-Lamas, P., T.M. Fernández-Caramés, M. Suárez-Albela, L. Castedo, and M. González-López. "A review on internet of things for defense and public safety." *Sensors (Basel, Switzerland)* 16, no. 10 (2016). https://doi.org/10.3390/s16101644.

Heath, D., D. Allum, and L. Dunckley. *Introductory Logic and Formal Methods*. Henley-on-Thames: Alfred Waller, 1994.

Hoare, C.A.R. "Notes on data structuring." In *Structured Programming*, edited by O.J. Dahl, E.W. Dijkstra, and C.A.R. Hoare, 83–174, London: Academic Press Ltd, 1972. http://dl.acm.org/citation.cfm?id=1243380.1243382.

International Telecommunications Union. "Minimum requirements related to technical performance for IMT-2020 radio interface(s)." 2017. https://www.itu.int/pub/R-REP-M.2410-2017.

Jacky, J. *The Way of Z: Practical Programming with Formal Methods*. Cambridge: Cambridge University Press, 1997.

Johnson, A.M., and S. Axinn. "The morality of autonomous robots." *Journal of Military Ethics* 12, no. 2 (2013): 129–141. https://doi.org/10.1080/15027570.2013.818399.

King, T.M., J. Arbon, D. Santiago, D. Adamo, W. Chin, and R. Shanmugam. "AI for testing today and tomorrow: Industry perspectives." In *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, 81–88, Newark, CA: IEEE, 2019. https://doi.org/10.1109/AITest.2019.000-3.

Kott, A., A. Swami, and B.J. West. "The internet of battle things." ArXiv:1712.08980 [Cs] (December 2017). http://arxiv.org/abs/1712.08980.

Lysaght, R.J., R. Harris, and W. Kelly. *Artificial Intelligence for Command and Control*. Willow Grove, PA: Analytics Inc, 1988. https://apps.dtic.mil/docs/citations/ADA229342.

McMahan, J. "Foreword." In *Who Should Die? The Ethics of Killing in War*, edited by R. Jenkins, M. Robillard, and B.J. Strawser, ix-xiv, Oxford: Oxford University Press, 2013.

Microsoft Defender ATP Research Team. "Protecting the protector: Hardening machine learning defenses against adversarial attacks." 2018. https://www.microsoft.com/security/blog/2018/08/09/protecting-the-protector-hardening-machine-learning-defenses-against-adversarial-attacks/.

Mirsky, Y., T. Mahler, I. Shelef, and Y. Elovici. "CT-GAN: Malicious tampering of 3D medical imagery using deep learning." *ResearchGate*, 2019. https://www.researchgate.net/publication/330357848_CT-GAN_Malicious_Tampering_of_3D_Medical_Imagery_using_Deep_Learning/figures?lo=1.

Moor, J.H. "What is computer ethics?" *Metaphilosophy* 16, no. 4 (1985): 266–275. https://doi.org/10.1111/j.1467-9973.1985.tb00173.x.

Morley, J., J. Cowls, M. Taddeo, and L. Floridi. "Ethical guidelines for COVID-19 tracing apps." *Nature* 582 (2020): 29–31.

Nagel, T. "War and massacre." *Philosophy and Public Affairs* 1 (1972): 123–144. In *American Behavioral Scientist* 15, no. 6 (1972): 951. https://doi.org/10.1177/000276427201500678.

O'Connell, M.E. "Banning autonomous killing: The legal and ethical requirement that humans make near-time lethal decisions." In *The American Way of Bombing: How Legal and Ethical Norms Change*, edited by M. Evangelista and H. Shue, 224–236, Ithaca: Cornell University Press, 2014.

Rigaki, M., and A. Elragal. "Adversarial deep learning against intrusion detection classifiers." *017 NATO IST-152 Workshop on Intelligent Autonomous Agents for Cyber Defence and Resilience,* 2017. https://ceur-ws.org/Vol-2057/Paper7.pdf.

Roberts, H., J. Cowls, J. Morley, M. Taddeo, V. Wang, and L. Floridi. "The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation." *AI & Society* (June 2020). https://doi.org/10.1007/s00146-020-00992-2.

Schubert, J., J. Brynielsson, M. Nilsson, and P. Svenmarck. "Artificial intelligence for decision support in command and control systems," *23rd International Command and Control Research & Technology Symposium "Multi-Domain C2"*. (2018).

Sharkey, A. "Autonomous weapons systems, killer robots and human dignity." *Ethics and Information Technology* 21, no. 2 (2019): 75–87. https://doi.org/10.1007/s10676-018-9494-0.

Sharkey, N. "Saying "No!" To lethal autonomous targeting." *Journal of Military Ethics* 9, no. 4 (2010): 369–383. https://doi.org/10.1080/15027570.2010.537903.

Sharkey, N. "Killing made easy: From joysticks to politics." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by P. Lin, K. Abney, and G. Bekey, 111–128, Cambridge, MA: MIT Press, 2012.

Sharkey, N.E. "The evitability of autonomous robot warfare." *International Review of the Red Cross* 94, no. 886 (2012): 787–799. https://doi.org/10.1017/S1816383112000732.

Sparrow, R. "Killer robots." *Journal of Applied Philosophy* 24, no. 1 (2007): 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Sparrow, R. "Robots and respect: Assessing the case against autonomous weapon systems." *Ethics & International Affairs* 30, no. 1 (2016): 93–116. https://doi.org/10.1017/S0892679415000647.

Taddeo, M. "Information warfare: A philosophical perspective." *Philosophy and Technology* 25, no. 1 (2012): 105–120.

Taddeo, M. "An analysis for a just cyber warfare." In *Fourth International Conference of Cyber Conflict*, Tallinn: NATO CCD COE and IEEE, 2012.

Taddeo, M. "Cyber security and individual rights, striking the right balance." *Philosophy & Technology* 26, no. 4 (2013): 353–356. https://doi.org/10.1007/s13347-013-0140-9.

Taddeo, M. "Just information warfare." *Topoi* (April 2014a): 1–12. https://doi.org/10.1007/s11245-014-9245-8.

Taddeo, M. "The struggle between liberties and authorities in the information age." *Science and Engineering Ethics* (September 2014b): 1–14. https://doi.org/10.1007/s11948-014-9586-0.

Taddeo, M. "On the risks of relying on analogies to understand cyber conflicts." *Minds and Machines* 26, no. 4 (2016): 317–321. https://doi.org/10.1007/s11023-016-9408-z.

Taddeo, M. "The limits of deterrence theory in cyberspace." *Philosophy & Technology* (2017a). https://doi.org/10.1007/s13347-017-0290-2.

Taddeo, M. "Cyber conflicts and political power in information societies." *Minds and Machines* 27, no. 2 (2017b): 265–268. https://doi.org/10.1007/s11023-017-9436-3.

Taddeo, M. "Trusting digital technologies correctly." *Minds and Machines* 27, no. 4 (2017c): 565–568. https://doi.org/10.1007/s11023-017-9450-5.

Taddeo, M. "Three ethical challenges of applications of artificial intelligence in cybersecurity." *Minds and Machines* 29, no. 2 (2019): 187–191. https://doi.org/10.1007/s11023-019-09504-8.

Taddeo, M., and L. Floridi. "The debate on the moral responsibilities of online service providers." *Science and Engineering Ethics* (November 2015). https://doi.org/10.1007/s11948-015-9734-1.

Taddeo, M., and L. Floridi. "Regulate artificial intelligence to avert cyber arms race." *Nature* 556, no. 7701 (2018): 296–298. https://doi.org/10.1038/d41586-018-04602-6.

Taddeo, M., T. McCutcheon, and L. Floridi. "Trusting artificial intelligence in cybersecurity is a double-edged sword." *Nature Machine Intelligence* 1, no. 12 (2019): 557–560. https://doi.org/10.1038/s42256-019-0109-1.

Taddeo, M., D. McNeish, A. Blanchard, and E. Edgar. "Ethical principles for zartificial intelligence in national defence." *Philosophy & Technology* 34 (2021): 1707–1729. https://doi.org/10.1007/s13347-021-00482-3.

Tamburrini, G. "On banning autonomous weapons systems: From deontological to wide consequential reasons." In *Autonomous Weapons Systems: Law, Ethics, Policy*, edited by B. Nehal, S. Beck, R. Geiβ, H.Y. Liu, and C. Kreβ, 122–142, Cambridge: Cambridge University Press, 2016.

"The UK and International Humanitarian Law 2018." n.d. Accessed 1 November 2020. https://www.gov.uk/government/publications/international-humanitarian-law-and-the-uk-government/uk-and-international-humanitarian-law-2018.

United Nations Office at Geneva. "Human Machine Touchpoints: The United Kingdom's Perspective on Human Control over Weapon Development and Targeting Cycles." 2018. https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/documents/GGE.2-WP1.pdf.

US Army. "Robotic and Autonomous Systems Strategy." 2017. https://www.tradoc.army.mil/Portals/14/Documents/RAS_Strategy.pdf.

World Economic Forum. "The global risks report 2018." *World Economic Forum*, 2018. http://www3.weforum.org/docs/WEF_GRR18_Report.pdf.

Yang, G.-Z., J. Bellingham, P.E. Dupont, P. Fischer, L. Floridi, R. Full, N. Jacobstein. "The grand challenges of science robotics." *Science Robotics* 3, no. 14 (2018): eaar7650. https://doi.org/10.1126/scirobotics.aar7650.

Zhuge, J., T. Holz, X. Han, C. Song, and W. Zou. "Collecting autonomous spreading malware using high-interaction honeypots." In *Information and Communications Security*, edited by S. Qing, H. Imai, and G. Wang, 438–451, Lecture Notes in Computer Science, Berlin: Springer, 2007.

8 Is Stuxnet the next Skynet? Autonomous cyber capabilities as lethal autonomous weapons systems

Louis Perez

In 2012, in his first public speech as the new leader of North Korea, Kim Jong-Un declared: "The days are gone forever when our enemies could blackmail us with nuclear bombs."[1] This proud assertion of the Korean leaders shows that having nuclear weapons is a strong strategic asset in international relations. Therefore, if you assume that one of your enemies is developing nuclear weapons, you will probably try to stop it. This is undoubtedly what the designers of the cyber worm Stuxnet had in mind in 2010, two years before Kim Jong-Un's speech, when they designed it to disable the centrifuges at the Natanz nuclear power plant in Iran.[2] Stuxnet has been recognized as one of the first cyber-attacks that caused physical damage[3] and therefore characterized as a cyber weapon.[4] Due to its independent functioning without human intervention and its automatic spread over the world, Stuxnet has also been labelled as an autonomous cyber weapon.[5] Autonomy is one of this ambiguous technological advance that inspire hope and fear. The fear of autonomous systems is particularly fuelled by science fiction and the rise of machines. This is the plot of the famous Terminator franchise. In these movies, the starting point is Skynet, an artificial intelligence (AI) system that has developed self-awareness and over which humans have lost control. Ironically, this AI system was designed to automate American nuclear response. Nevertheless, the public did not retain the idea of the AI system but rather their physical envelope, the autonomous robots known as the Terminator. Wondering if autonomous cyber capabilities (ACC), such as Stuxnet, may soon become as the pure fictional Skynet system, the answer is no, considering current and near-term technologies.[6] However, it draws attention to how autonomy is increasingly being used in cyber capabilities and how States intend to manage this in light of potential threats, especially in warfare.

Autonomy and cyber, and their potential uses in warfare, are two areas of concerns for States. Thus, they met regularly in two distinct international forums to discuss these topics. As early as 1998, cyber, referred to as information and communications technologies, is a topic of discussion within the UN First Committee.[7] In 2004, a Group of Governmental Experts on Advancing responsible State behaviour in cyberspace in the context of international security (UN cyber GGE) was established with 15 members.[8]

DOI: 10.4324/9781003284093-11

However, it failed to reach a consensus and issue a report. Nevertheless, reports were subsequently delivered after the creation of others GGE in 2010,[9] 2013,[10] 2015.[11] After another lack of consensus at the 2017 GGE, a new GGE was created in 2019 and released a report in 2021.[12] In parallel with the GGE, an open-ended working group (OEWG), open to all UN Member States, was established. The OEWG released its final report on March 2021.[13]

Autonomy is also a topic in a very specific forum, the Meeting of High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons (CCW).[14] In 2013, States parties decided to convene an informal Meeting of Experts to discuss the questions related to emerging technologies in the area of lethal autonomous weapons systems (LAWS). After three years of informal meetings, States decided to establish a Group of Governmental Experts (CCW GGE) on LAWS in 2016.[15] After three years of discussions, States successfully agreed on 11 guiding principles on LAWS in 2019.[16] Principles especially address issues such as the application of international law, notably international humanitarian law (IHL), weapon reviews, responsibility and accountability, human control, etc.

States remain relatively silent on the interaction between autonomy and cyber and how international law should apply to it. Indeed cyber and LAWS are seen as two different subjects that do not cross. This is even more surprising considering the proximity between the both. Indeed autonomous and cyber systems have been described as both software-based.[17] The United States Department of Defense (DoD) science board described cyber as "broadly used to address all digital automation used by the Department and its industrial base. This includes weapons systems and their platforms."[18] Therefore, what is said about LAWS could, in part, apply to cyber systems. In fact, this software base is not ignored from the CCW GGE. One of the aforementioned guiding principles acknowledges that LAWS are software-based by calling on States to consider "cyber-security against hacking or data spoofing."[19] However, it seems that code has been personified in relation to LAWS. Although software-based, LAWS are considered physical, whereas cyber systems are considered immaterial, despite their physical roots.[20] The idea of an interplay between autonomous weapon systems and cyber systems appears incongruous for States. Yet, beyond the Stuxnet case, cyber systems seem to benefit more and more from autonomous capabilities, especially thanks to AI algorithms.[21]

Autonomy promises to enhance system capabilities such as speed and accuracy in a way that exceed any human capabilities in this area.[22] Hence, in cyber warfare, autonomous systems will be a great advantage, especially to execute rapidly mundane tasks but also more complex ones in changing environments.[23] There is an increasing need of adaptative cyber systems regarding the necessary work required to ensure adequate defence.[24] Nevertheless, clarifications are needed to assess what is an autonomous system. To a certain extent, it has been asserted that because cyber systems rely on software and conducted on computers, automation is always present.[25] However one may wonder to what extent automated is part of autonomy and if there is

difference from simple automated cyber scripts and very complex ones, such as Stuxnet, that are often characterized as autonomous.[26] It is unclear if autonomy only refers to "fully autonomous" cyber systems or if advanced automated system could in a sense also be view as autonomous. Fully autonomous systems that could activate and (re)program itself and replace cyber human specialist are far from existing yet[27] and Stuxnet was clearly not one of them.[28] One of the reasons is that great part of cyber activities still need humans and can't be automated, although more and more are.[29] Given the increasingly automation of cyber capabilities and the uncertain point that distinguish automation from autonomy, authors have suggested broader definitions of autonomy. An autonomous system has been described as a system that "generates the rules by which it operates in its environment, and that no other entity generates the rules by which the system operates"[30] or the ability of a system "to perform some task without requiring real-time interaction with a human operator."[31] Several techniques are used to enhance autonomous features of cyber systems, especially AI algorithms that allow the system to be self-adaptive[32] and self-learning.[33] Therefore, offensive and defensive ACC are and will be used in conflict.[34] Several examples such as the DARPA Cyber Grand Challenge demonstrate these capabilities.[35] ACC can be used in several steps of an attack such as reconnaissance, infiltration and command and control.[36]

However, the exact scope of what autonomy is and how concerns related to LAWS, including international law, apply to ACC are open questions. As a result, the issue of ACC seems to elude both forums despite their current existence and foreseeable developments, notably in warfare. Hence, the present chapter will explore an innovative way to apply international law to these capabilities by characterizing them as LAWS and subsequently applying the CCW GGE outcomes to such systems. The scope of the chapter will mainly focus on States positions within CCW debates on LAWS and how these positions could apply to cyber means. The analysis will be conducted from the various diplomatic positions and interpretations of international law expressed by States regarding LAWS definition and IHL compliance. The purpose of such an approach is to highlight the advantages and adaptations needed for the characterization of autonomous cyber means. For instance, what the specific environment of cyberspace would change with regards of CCW discussions about LAWS. Moreover, this approach will involve applying States' positions on cyber means in a forum other than the UN GGE and OEWG. An useful dialogue will necessarily occur between the States' interpretations regarding LAWS and cyber. Therefore, the chapter will demonstrate how the States' positions on ACC as LAWS may impact their position on every cyber system.

In the first part, an analysis of States debates related to the LAWS definition will be conducted. Examining States position on the definition of LAWS, and on each word composing the expression, it will be concluded that according to some States' interpretations within the CCW,

ACC could be characterized as LAWS. From this conclusion, another part will explore how current debates on IHL regarding physical LAWS would apply to such systems in cyberspace. This part will especially focus on weapon reviews, human control and IHL compliance of cyber autonomous capabilities as LAWS.

Autonomous cyber capabilities defined as lethal autonomous weapon systems

This section will describe how some ACC may be characterized as LAWS. First, an analysis of States' discussions on the definition of LAWS will be conducted. Noting that almost no State has ever referred to ACC as LAWS, it will be examined how States interpretations on each word and concept related to LAWS might include cyber means and ACC. The section will conclude that, based on some interpretations of the core concepts of LAWS, ACC could fit a definition of LAWS.

The state of the debate: have autonomous cyber capabilities ever been defined as lethal autonomous weapon systems?

The interaction between cyber and autonomy is known from the various actors who are involved in these subjects. ACC are sparingly referred to as LAWS.[37] However, several concerns have been raised about the potential threats posed by the interplay between cyber and autonomy. In 2021, in the Chair's Summary of the OEWG, it has been pointed out that "pursuit of increasing automation and autonomy in ICT operations was put forward as a specific concern."[38] While some States within the OEWG recognize the increasing importance of autonomy in cyber systems, they do not refer to them as potential autonomous weapons.

In 2012, the United States DoD adopted the Directive 3000.09 addressing autonomy in weapon systems.[39] This directive provides a definition of autonomous weapon systems, however it states that this definition does not apply to certain systems, including "autonomous or semi-autonomous cyberspace systems for cyberspace operations."[40] Two elements must be specified regarding this directive. First, it is an undeniable recognition that autonomous and semi-autonomous cyber systems exist and are used in cyber operations. Second, omitting autonomous cyber systems as part of autonomous weapons should not be understood as a rejection from the United States of autonomous cyber weapons and was essentially justified on practical grounds. It was argued that making a precise classification between these types of autonomous systems would have taken a long time while the directive was fast needed to establish the US position on the issue of autonomous weapons.[41]

Therefore within this international forum and this internal policy, some States, acknowledge the interplay between cyber and autonomy and the potential uses in conflict but never directly as autonomous weapons. Nevertheless,

within the CCW GGE the idea of cyber as AWS emerged. Indeed, experts underlined that autonomous weapons will probably be first developed in cyberspace.[42] They also highlighted the lack of discussion about the fact that cyber systems could be autonomous weapons and that cyber weapons also lacked human control.[43] Thus, they urged States to consider these points, especially if these cyberoperations have kinetic effects.[44]

Beyond these experts views, there is one case where a State appears to have recognized ACC as LAWS. Indeed, in its commentaries on the operationalization of the CCW GGE guiding principles at the national level, Portugal stated:

> Indeed, a hostile actor — be it a State actor, a non-State actor or an actor by proxy — in the possession of LAWS may use them as an asymmetric tool/mean of warfare or of cyberthreat. For example, LAWS could be used for espionage actions, intrusion/infiltration activities or in attacks against persons, facilities or networks located abroad, in violation of international law and in an undercover fashion making detection and accountability difficult or even impossible.[45]

In this statement, Portugal recognized that cyber means could be seen as LAWS and used to perpetrate attacks against persons and objects. It is the unique occurrence found within the GGE debates where a State directly discussed LAWS as cyber means in cyberspace with potential kinetics effects on persons or objects. It should be noted that Portugal does not justify this point but rather admit it as an obviousness, suggesting that ACC are inherent to LAWS. When using the expression LAWS, it could indistinctly refers to physical or cyber systems. This view seems somewhat extrapolated giving the fact that, apart Portugal, no State has ever defined autonomous cyber systems as LAWS. However, this raises the question of whether a definition of LAWS could include autonomous cyber systems.

The debate of the States: the definition of lethal autonomous weapon systems and the potential inclusion of autonomous cyber capabilities

Defining cyber capabilities as LAWS is not an easy task given the fact that there is no consensus among States on what LAWS are even though numerous States submitted their definitions of LAWS.[46] One may argue this is unprecedented in the history arm control treaty negotiations where the definition of weapons is not the core of the debate and often come at the end process. Within the GGE on LAWS, the definition and the necessity of a definition is a debate on itself. This section will consider the stake of the debates and analyse, from the States perspective, each term comprising LAWS ("lethal", "autonomous" "weapons systems") in order to assess if cyber capabilities may be understood as part of LAWS.

The debate on the necessity of a definition of LAWS

Numerous States recognize that a definition of LAWS could be useful,[47] and declared it was vital to address potential risks that LAWS could pose.[48] One argument related to the necessity of a definition is that different definitions of LAWS may lead to confusion and misunderstanding.[49] On the other hand, some States have suggested that having a current definition was void of meaning, maintaining that LAWS do not exist and therefore can't be defined.[50] It has also been argued that given the fact that the area of LAWS is evolutive and involved dual-use technologies it would be too difficult to agree on a common and constant definition.[51]

The purpose of the definition and its link to a regulation was also highlighted. A State indicated that the definition debate "turned into a political issue suiting the respective policy positions."[52] Another State underlined that "a legal definition is generally developed for the specific purposes of a legal rule and not in the abstract. Often legal definitions determine the scope of a legal rule."[53] Thus, in order to define LAWS, States must determine what are the relevant legal rules and to which systems they apply. Concurrently, States recalled that the purpose of defining LAWS should be viewed independently from the regulation[54] and should not consist in designing what are the "good" and "bad" systems but rather to identify the types of systems the GGE should address.[55]

Therefore, considering the current debate and the progress achieved on many related issues since 2014, without a comprehensive working definition,[56] several States argued that, so far, a definition was not necessary for the continuation of the work.[57] Consequently, States agreed upon an approach focusing on characteristics and concepts related to LAWS rather than a common definition.[58] This approach would allow for a common understanding of the issues, and avoiding the use of the same words with different meaning and overall flexibility with regard to the evolutive nature of new technologies.[59]

This position is illustrated by the current mandate of the GGE which refers to "emerging technologies in the area of LAWS."[60] Interestingly, the mandate of the previous informal group was only about LAWS. Following the characteristics approach, it has been decided to expand the scope of the GGE mandate to emerging technologies rather than solely LAWS which remains undefined. Therefore within this forum, States should address all emerging technologies that contribute to LAWS. But, as South Africa pointed out, there is no definition of emerging technologies which remains an open-ended concept.[61] Considering this concept, several States specified what this could refer. Some highlights it was notably AI[62] whereas others indicate autonomy.[63] The United States recalled that, according to its own definition of the term, autonomy in weapon systems is not new and has been used for decades.[64] Significantly, no State mentioned cyber as an emerging technology in the area of LAWS despite the common software based discussed above. One may quickly close the debate of ACC as LAWS, at least as subject to the GGE

CCW discussions, by considering cyber means as such emerging technologies and therefore including them within GGE mandate.

The analysis of the lethal autonomous weapon systems characteristics from a states' perspective

This section will analyse the different meanings, according to States, given to the terms which form "LAWS." The purpose is to determine if the ACC could be part of the scope of each element and result in a specific definition that incorporates ACC in LAWS.

"Weapon systems"

Weapons are at the core of the CCW purposes. The CCW was designed in order to regulate and prohibit certain weapons in accordance with IHL rules.[65] However, this convention does not define the term "weapons", focusing instead on the effects of the use of a weapon rather than weapon as the tool itself. The advantage of this approach is that it provides a flexible scope to include new forms of weapons, such as "cyber weapons" and "autonomous" weapons.

"Weapon systems" is the term used in the name and the mandate of the CCW GGE. In the quest of defining LAWS, or at least their characteristics, States tried to understand the exact meaning a "weapon systems." In a working paper on the weapons review mechanisms, Netherlands and Switzerland estimated that "from a tactical point of view, almost anything can be used as a weapon."[66] They state that a

> 'weapon' is not limited to weapons in the traditional sense (firearms, artillery pieces, etc.) but includes all objects, devices, etc., […] provided that those objects, etc., are intended to cause harm to persons (including injury or death) or damage to objects (including destruction, capture and neutralization) (…).[67]

Other States support this assertion. Estonia and Finland, highlighted that weapons are not just projectiles but also "other capabilities, such as lasers, high power microwave (HPM), nanoparticles, or other mechanisms, [that] could potentially be used to cause harm to an adversary."[68] Another State indicates that a weapon system "refers to a weapon along with any associated technology necessary for the operation of the weapon."[69] These statements demonstrate that a projectile, or even a physical object, is not necessary to characterize a weapon. This open a space for including cyber means as weapons within the CCW discussion. Indeed, it is recognized that what is important is the way and the intent which the named weapon will be used. According to this approach, as long as cyber capabilities would result in harm or damage, it could be qualified as weapons. Nonetheless,

with the exception of Portugal, no State has mentioned cyber capabilities as LAWS or weapons.

Considering this question, several experts recognize cyber means could be characterized as weapons. In particular, a rule of the Tallinn Manual 2.0 characterizes cyber weapons as means of warfare.[70] The commentaries precise that cyber weapons are "used to cause injury to, or death of, persons or damage to, or destruction of, objects, that is, that result in the consequences required for qualification of a cyber operation as an attack" and "include any cyber device, materiel, instrument, mechanism, equipment, or software used, designed, or intended to be used to conduct a cyber-attack."[71] Despite this manual does not reflect States positions, it is consistent with the previous positions mentioned: a weapon is the tool used to conduct an attack, which causes harm or damage.

It is interesting to note that in the UN cyber forums, there is no agreement on whether cyber means could be characterized as weapons as well. The matter is not mentioned in the various reports of the cyber GGE and of the OEWG. Because the UN cyber GGE process is quite opaque, there is few information from this forum. However, in a 2019, France used the expression cyber weapons through its statement.[72] Conversely, the OEWG is more transparent. In this forum many States refer to weaponization of cyber means[73] while some directly used the expression cyber weapons.[74] Some States, such as the United Kingdom and Cuba, seem to oppose cyber means to conventional weapons.[75] Eventually, Russia clearly refuses to refer to cyber means as weapons.[76]

Recognition of ACC as weapons in the framework of the CCW could contribute to broader recognition of cyber means as weapons. Indeed, if the CCW GGE considers cyber capabilities as LAWS, or at least if it accepts a large definition that possibly includes cyber capabilities, this could be a strong argument for accepting certain cyber means as weapons in other forums. However, States opposed to this idea in UN cyber discussions are unlikely to let this happen within the CCW GGE.

From this discussion one may conclude that certain ACC may be characterized as weapons due to potential harmful consequences. However, harms or damages could not suffice to qualify such systems as LAWS as it seems to require a necessarily lethal effect.

"Lethal"

The "weapon systems" mentioned in the CCW GGE mandate are intended to be lethal. Even though the word "lethal" is part of the expression, questions have been raised regarding the necessity of this lethal element to characterize LAWS. Several States wondered if AWS should inevitably have to be lethal.[77]

On the one hand, some States understand lethality as basic and crucial characteristics of LAWS beside autonomy. For example, Cuba states "The

greater the autonomy and lethality, the stricter the framework that regulates them should be."[78] Thus, it has been suggested that the discussion limits only to lethal AWS.[79] One argument raised in favour of this opinion relied on the fact that the

> Inclusion of non-lethal autonomous weapons systems, as a part of LAWS, would unnecessarily expand the scope of rules too broadly, which could end up driving States to evade such rules for the sake of their own national security, and accordingly lead to a loss of effectiveness of the rules.[80]

According to these statements, the lethality of a weapon is an underlying condition of LAWS, particularly to ensure the effectiveness of their regulation. A State specified that lethality means the weapon has a "sufficient payload and for means to be lethal."[81]

On the other hand, numerous States have taken the opposite side considering that lethality "should not be conceptually regarded as a prerequisite characteristic of autonomous weapon systems."[82] They argue that if lethality is retained as an element of AWS, it would be "a novel concept in the CCW framework" that was not a prerequisite mentioned during negotiations related to blinding lasers and non-detectable fragments.[83] Various references to IHL were also made underlining that "lethality is not a defining feature of any weapon system."[84] Indeed, one State argued that "lethality *per se* is not a concept in IHL" and that "lethal weapons may be used in compliance with IHL."[85] Conversely, another State declare that non-lethal weapon may have "lethal impacts in certain circumstances but where the lethal effect is not the primary purpose of the system."[86] Therefore, a non-lethal weapon may contravene IHL rules.[87] Eventually, States criticize that a focus on lethally would elude damages to certain goods that are also protected by IHL.

Regarding this debate, Switzerland recommended to have a broad understanding of AWS

> which would also cover means and methods of warfare that do not necessarily inflict physical death, but the effects of which may be restricted to causing, for example: (1) physical injury short of death, (2) physical destruction of objects, or (3) non-kinetic effects.[88]

It also recommended to adopt an IHL-centred approach and to focus on "attack" that is a concept defined in the Additional Protocol I.[89] An interesting parallel should be drown with the previous discussion on the definition of a weapon. Again, for some States, the landmark should be on systems that launch an attack, regardless of whether the system is already qualified as weapons or necessarily lethal. Interestingly, Switzerland also mentioned non kinetic effect which clearly include cyber dimension.[90]

Wondering if a cyber weapon may lead to death, the answer is yes according the experts of the Tallinn Manual.[91] However, the scope is extremely

limited if only cyber LAWS are subject to the CCW GGE mandate. This will affect very few systems and probably no current one. Therefore, if one want to address effectively ACC as AWS, it is preferable to avoid such lethal requirement.

To conclude, the term "lethality" implies a focus on the effects of AWS. Whether lethality is part of the final definition of LAWS or not, ACC may be included in such definition. The sole difference will be the number of systems that fall within the scope of this definition since fewer cyber systems can result in lethality and, so far, none are inherently lethal.

"Autonomy"

As we mentioned above, autonomy in cyber capabilities has been asserted by some experts. Autonomy is an important element of the CCW discussions given that it is often considered as the problematic characteristic that make weapon systems potentially against IHL.

Within the CCW, the definition of autonomy has been the subject of much discussion and has brought out opposing views. From this debate it will be determined whether ACC correspond in part to the conclusion of this debate.

One of the main conceptions of autonomy within the GGE discussions address autonomy as a spectrum[92] that varies "from zero to full autonomy."[93] The categories following this conception often strictly distinguish system as automated, semi-autonomous or fully-autonomous. Many States considers that LAWS are not automated or remotely operated weapon systems.[94] Thereby, some of these States views LAWS as only fully autonomous systems.[95] States indicate that fully AWS does not exist yet[96] and may never exist.[97] States, such as Cuba[98] and Russia,[99] suggested definition of fully autonomous and semi-autonomous weapons.[100] However, many States highlight that there is no clear reference point for when a system becomes fully autonomous[101] and that a definition of LAWS should provide the point in the scale.[102] From this, one can infer the intention to differentiate autonomous systems from other systems. However, the determination of what constitutes an autonomous system and how it contrasts with other systems is far from a consensus.

Due to this unclarity, States criticized the use of the word autonomy as an "on/off" phenomenon[103] and underlined it was not a "binary technology."[104] It has also been noted that the degree of autonomy "can go back and forth – from manual, to semi-autonomous, to autonomous – and just because an autonomous system has a manual button does not mean it is not to be considered."[105] Hence, there may be no single general level of autonomy across a system.[106] A State affirmed that autonomy "is a relative term, with different understandings in different disciplines"[107] and denigrated the use of the expression fully AWS, stating that it seems to imply that there is a clear line, rather than a continuum.[108] Another State added that the expression "fully autonomous" was imprecise and potentially unhelpful.[109] From this part of

the debates, it can be concluded that cyber AWS might have different degree of autonomy and should not be considered as fully autonomous systems, although the definition of the latter may vary.

Without rejecting the spectrum conception, many States insisted that what matters is not autonomy *per se* or the autonomy of the system, but rather autonomy in function or task and the human involvement in the execution of the task.[110] Numerous States indicate that it is primarily autonomy in critical functions, notably in the targeting cycle, that is of concerns.[111] Critical functions were identified as a good way to overcome issues of defining autonomy.[112] It was stated it was useful to differentiate this kind of AWS from others.[113] Some States argue that autonomy in critical function already exists in weapon systems.[114] Another important element raised in assessing the autonomy in function, or in the system is the capacity of self-learning and self-evolution, especially in critical functions.[115] Such autonomous functionality in the operation of a system has been designated of particular concern with respect to human control, even though other autonomous systems without this functionality remains concerning.[116]

Beyond this technical perspective, many States recall that autonomy should also be understood as an interaction between human and machine.[117] It has been suggested that this approach will allow to go beyond a technological prism that is dependent on the evolution of technology and will ensure the implementation of legal obligation.[118] For this reason, a number of States have called for maintaining a human control over critical functions.[119] This position is consistent with the above-mentioned experts' views of autonomous cyber systems that are referring to the human involvement.

In sum two main elements seems to emerge from this brief summary of the CCW debates concerning autonomy. These two elements are the two faces of the autonomy coin:[120] the technological face address the technical autonomy of systems that can interact by themselves with a changing environment and operate in uncertainty, whereas the human interaction face address the control humans have over autonomous systems. However, multiple States interpretations and definitions may result from these elements and, so far, there is not a clear definition, or even boundaries, regarding the concept of autonomy within the CCW.[121]

Because the debate is on the concept of autonomy, applying this debate to "autonomous" cyber capabilities implies the same issues: is there a spectrum from automated to fully ACC or is it more relative? What cyber critical functions can be autonomous? What control human exerts on such cyber systems or functions?

The Stuxnet case illustrates the various possibilities of characterizing a cyber system as autonomous or not. Many authors consider that Stuxnet had a certain degree of autonomy.[122] Indeed, due to the specific circumstances, Stuxnet acted on the local network which was not connected to internet and therefore without any direct and external human intervention. It executed

critical functions (propagate, identify, infect and control the target, self-update) by itself to achieved its assigned objective. Therefore, the absence of human control as well as its ability to interact with its environment lead to characterize Stuxnet as an autonomous system. This lack of human control is consistent with the previously mentioned example of fully autonomous system definitions provided by some States. Nevertheless, some authors underlined that Stuxnet was not autonomous but rather automated.[123] Indeed, Stuxnet did not act autonomously insofar it merely followed steps programmed into its code to execute human determined objectives. Once inside the local system, the worm surely acted without human control, but human reconnaissance and human computing expertise were required to design such system.[124] Therefore, this demonstrates that depending on the interpretation of human control, the vision of the autonomy of a system might vary. If Stuxnet had capabilities as self-learning or self-evolving, these authors might have characterized it as autonomous but it appears that the execution of critical functions without direct human control was insufficient to see it as autonomous. One may argue that even if Stuxnet was automated, it made mistakes and produced unexpected results, such as the propagation of the worm outside Iran.[125] These unexpected results demonstrates that such absence of human control is worrisome, even if the system is "only" automated. More broadly, discussions surrounding the characterization of Stuxnet as automated or autonomous illustrates the limits of this semantic debate. The novelty, complexity and variety of tasks executed by Stuxnet have created confusion on its nature and its potential autonomous features. However, whether the system is autonomous or automated, it may lead to critical results and unlawful actions. Therefore, the way the CCW parties will define autonomy will be crucial because it might exclude automated systems that could have unlawful consequences.

In any case, regarding the characterization of cyber capabilities as LAWS, whatever the definition of autonomy adopted by the CCW, some cyber systems will always seem capable to fit such definition due to the software based of such systems. If the autonomy is somewhere in a system or a function, it will partly come from the software and therefore linked to cyber characteristics. Thus, this debate does not preclude to characterize ACC as autonomous in the sense that some States understand and interpret autonomy within the framework of the CCW.

From all of the above, the characterization of ACC remains an open possibility. Indeed, due to the variety of interpretations for each concepts related to LAWS, there is space where cyber capabilities can fit. It is interesting to note that for each concept, a bottom-up approach was suggested to avoid definition issues. Following this potential characterization of cyber means as AWS, the next part of the chapter will analyse what are the implications of applying the GGE CCW outcomes to such systems with regard to IHL compliance.

Cyber autonomous weapon systems and international humanitarian law

Considering ACC as AWS, also referred to hereafter as cyber AWS, implies that such capabilities are subject to the CCW GGE mandate and therefore affected by the discussions. Many topics are at the forefront of the discussions within the forum: mainly IHL, but also responsibility and accountability, ethics, international security issues, etc. Thus, the choice has been made to focus on IHL in this chapter. This choice relies on several reasons. The inextricable link between the CCW and IHL encourages to primarily focus on it. Indeed, a great part of the debates address IHL issues, such as applicability or weapon review. In addition, dealing with IHL issues can help anticipate other issues. For instance, ensuring IHL compliance would have consequences on responsibility and ethics. Therefore, the next sections will address the applicability and the application of IHL to ACC in light of the CCW's GGE discussions.

The applicability of international humanitarian law to cyber autonomous weapon systems

Not surprisingly, the GGE recognizes the applicability of IHL to LAWS in its first principle.[126] Indeed, CCW is based on IHL and its material scope relies on Geneva Conventions.[127] Thereby, admitting certain ACC as LAWS will conduct to fully recognize that IHL applies to some cyber means of warfare.

However, the IHL applicability to cyber means of warfare has not always been explicit and is still a source of discussions. NATO,[128] European Union[129] and numerous States[130] recognized the applicability of IHL to cyber operations during armed conflicts. This has also been demonstrated by experts of the Tallinn Manual.[131] Notwithstanding, several States refused to recognize such applicability. Within the UN cyber GGE the applicability of international law and especially IHL to cyberspace is a recurrent matter. A major step was taken in the 2015 report of the UN cyber GGE where States, "note[d] established international legal principles including, where applicable the principles of humanity, necessity, proportionality and distinction."[132] Without naming it, States recognized that the core principles of IHL apply in some, unspecified, circumstances to cyberspace.[133]

However, difficulties tarnishing this success emerged in the discussions that followed. Despite the discreet reconnaissance of IHL's applicability in the 2015 report, several States, like China, Cuba or Russia, objected to use the term of IHL. This position mainly relied on the argument that the applicability of IHL "would legitimize a scenario of war and military actions in the context of ICT."[134] This contributed to the failure of the UN cyber GGE to adopt a report in 2016–2017. Interestingly, the same type of argument was used within the CCW discussions. The report of an informal meeting of

experts on LAWS indicates that "several delegations cautioned against taking for granted that existing IHL applies to LAWS and that by doing so, these weapons could be prematurely legitimized."[135] Again, several States seem to view the applicability of IHL as a potential threat due to legitimization such means in wartime. The legal (ir)relevance of such arguments, which seems more political than legal,[136] will not be discussed here. Thereafter, the OEWG Chair's report highlighted that States recalled IHL "neither encourages militarization nor legitimizes resort to conflict in any domain."[137]

This undeniable applicability of IHL to LAWS, and in this paper to cyber AWS, could push States to recognize the applicability to other cyber means of warfare. Indeed, the recognition of the applicability of IHL to cyber AWS, implies a recognition of the applicability of IHL in cyberspace is those specific circumstances. This would demonstrate the inconsistency of the political argument mentioned above. In that sense, efforts have been made recently. In 2021, the report adopted by the UN cyber GGE indicates that "international humanitarian law applies only in situations of armed conflict."[138] While IHL is finally mentioned, it is recognized as applicable "only" in armed conflict. Even though IHL is the law of armed conflicts, several IHL obligations apply in peacetime such as the legal review of new means and methods of warfare. Hence, the wording adopted by the GGE may be confusing regarding the scope of IHL rules that apply to cyber means. Again, the work of the CCW GGE is useful in this regard. States have recognized that the compliance with IHL is not limited to the rules related to the conduct of hostilities[139] and that peacetime IHL rules should be taken into account.[140] The application of the CCW GGE discussions regarding the applicability of IHL rules to cyber AWS is therefore helpful to support the applicability of all the relevant IHL rules to cyberspace in these circumstances. A State which admits this applicability of IHL to LAWS could not deny it to any cyber means of warfare because LAWS would include cyber AWS.

The application of international humanitarian law to cyber autonomous weapon systems

If the applicability of IHL could seem a less striking issue, the question of how IHL should apply is a major one. Whether for cyber operations or LAWS, States had discussions about how IHL rules should apply. Regarding cyber, the Tallinn Manual brings several answers and interpretations of IHL rules. However this document only express experts' views. States may have different opinions. Thereon, the last GGE requested States, on a voluntary basis to share their understanding of how international law applies. The report was released in July 2021.[141]

IHL application is also one of the main subjects within the CCW. Main concerns are related to the obligation to conduct a weapon review and to assess the way LAWS comply or not with IHL principles. These two subjects will be analyse from a cyber AWS perspective.

The legal review of cyber autonomous weapon systems

According to the article 36 of AP I,

> in the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.[142]

In sum, a State has the obligation to assess the compliance of a new weapon to international law, especially IHL. This article is not considered as part of customary international law and therefore only applies to the 174 States parties to AP I.[143] However, customary international law still requires from every State to ensure the means of warfare used complies with IHL rules, whether customary rules or treaty rules.[144]

In this part, the CCW GGE weapon review mechanism discussions will be mainly analysed with regards to its potential contributions to international law and cyber discussions. Thus, it will not examine in depth the conditions and content of such legal review regarding cyber AWS. Nonetheless, the main question of cyber AWS compliance with IHL will be address in the next section.

The weapon review debates and cyber capabilities

One of the major merits of CCW GGE discussions has been to highlight the weapon review mechanism, notably mentioned in the guiding principle (e).[145] This spotlight on the legal review requirement brought a lot of debates on what exactly is this review and how to apply it to LAWS. Indeed, characteristics of LAWS such as autonomy and self-learning capacity result in a potential unpredictably which force States to ensure the proper functioning of the system through such a weapon review.[146]

Despite the usefulness of the weapon review in ensuring IHL compliance, several States argue that this obligation relies on national procedures that will result to a lack of international uniformity and different standards.[147] To address this problem, many States suggested that there should be a voluntary exchange of best practices on the conduct of the legal reviews on emerging technologies in the area of LAWS.[148] Argentina particularly emphasized that this sharing could be based on article 84 of the PA I which encourages such an exchange of information.[149] This exchange was described as a confidence-building measure that could enhance transparency and trust in the use of LAWS and compliance with IHL.[150] It was proposed that this process relied on a specific mechanism[151] where States would share information on their national implementation of the weapon review.[152] This exchange could then lead to a comparative analysis.[153] Demonstrations of LAWS have even been suggested.[154] However, States recalled there was no obligation

to share information about the review and its results.[155] The limits of national security and industrial property rights in sharing information were also raised.[156] More pessimistically, Austria indicates information was "unlikely to be shared in real time with the broader international community"[157] and other States pointed out it will be "difficult to assess the quality of weapon reviews given only a limited number are publicly available."[158] Notwithstanding, several States already shared information on their weapons reviews like Australia,[159] Germany,[160] Netherlands,[161] Russia,[162] Sweden[163] or the United Kingdom.[164] Considering this dynamic and with regards to cyber AWS, one may assume that States will be encouraged to share information on the specific case of ACC if there are defined as AWS. In addition, these information exchanges, referred to as confidence-building measures, can enhance those already discussed in UN cyber forums.[165]

It's noteworthy that even if States refuse to define cyber AWS as LAWS, the current debates within the CCW remind to States their duty to review new weapons, means and methods of warfare. Therefore, the work and debates on how to assess autonomous systems such as LAWS could prove to be useful for States to evaluate autonomous and non-autonomous cyber means of warfare. Even though the experts of the Tallinn Manual provide a constructive analysis of the weapons review to cyber means of warfare,[166] States remained relatively silent on that matter which is hardly discussed within UN cyber forums.[167] In 2021, in accordance with the call of the cyber GGE to States to share their national views on how international law applies in cyberspace, Australia recognized that "[a] cyber capability could, in certain circumstances, constitute a 'weapon, or a means or method of warfare' within the meaning of Article 36 and require a review."[168] One may suppose this a result of the CCW GGE dynamic as Australia demonstrated to be very active on article 36 matter.[169] Other States such as Brazil and Switzerland also recognized the necessity of article 36 to cyber means.[170]

Furthermore, recalling the discussion with the expression "emerging technologies in the area of LAWS," one may consider that cyber is one of these emerging technologies contributing to LAWS. This is obvious if LAWS include cyber AWS, as in this chapter, but it appears less obvious for physical AWS as their software base is often omitted. Thus cyber capabilities used in AWS, whether cyber or physical, should be assessed in the weapons review.

Perspectives on the application of the weapon review to cyber autonomous weapon systems

The cyber dimension of cyber AWS will impact the legal review. Such a review promises to be very difficult because it combines two technologies for which there are already many issues.[171] This would require an entire chapter, thus the present one will just share some perspectives on that topic here.[172] Addressing the question of who should conduct such review States within the CCW GGE debates required this should be done by a multidisciplinary team.[173] Considering cyber AWS, the weapon review team should include

cyber and AI experts beside the legal ones. On the elements that needs to be reviewed regarding AWS, specific computer hardware and software, data, environment of use have to be assessed. This may be challenging given the fact that cyber data and environment are very different from physical ones. With respect to the timing the review should take place, cyber AWS reviewer will have to deal with the recurrent updates of the cyber AWS software, probably more than physical ones. The self-learning and self-adaptative algorithms that may be used in cyber AWS are also of concerns.[174] For instance, if a cyber AWS changes its own code, it should probably be reviewed again.[175] Therefore, a cyber AWS acting without direct human intervention, like Stuxnet, should not have such capabilities because a review would be required but impossible.[176]

Human control over cyber autonomous weapon systems and international humanitarian law compliance

Interestingly, human control has rarely been mentioned with respect to cyber systems.[177] Conversely, it is one of the core topics within the CCW GGE, notably addressed by the guiding principle (c).[178] Human control raises the question of the extent of human-machine interaction required to comply with IHL. There are two opposing positions on this subject.[179] On the one hand, following a technological approach, States argue that human control relied on technical characteristics. If IHL assessments can be programmed in a AWS, there is no problem, especially considering the alleged superiority of machines over humans. This position, however, does not exclude human from any control but adapt humans involvement regarding technical means. On the other hand, following a more humanistic approach, States believe that IHL assessments require human judgment in all cases.[180] While technologies can help humans comply with IHL, machines can't enforce IHL rules. This debate exceeds the LAWS discussions and applies to technology in general. With respect to cyber warfare, it seems that the first view is commonly shared. At least it seems rarely argued that a cyber system requires human judgement in its conduct.[181] The difference probably lies in the context of utilization. In war time, this position may change especially considering ACC. The present section will explore the relation between cyber AWS and IHL.

Cyber AWS and the enhancement of IHL

Fear of a loss of control of AWS and potential subsequent violations to IHL rules usually surround the CCW debates. Against the flow, some States, supporting the technological approach, point out that AWS could also be a better way to ensure IHL. This position is reflected in guiding principle (h) stating "consideration should be given to the use of emerging technologies in the area of lethal autonomous weapons systems in upholding compliance with IHL and other applicable international legal obligations."[182]

In 2018, the United States provided a working paper on "Humanitarian benefits of emerging technologies in the area of lethal autonomous weapon systems."[183] Throughout the document the State describes functions and examples of ways in which LAWS could reduce risks to civilians such as autonomous self-destruct, self-deactivation or self-neutralization mechanisms, increase of military awareness due to analysis of data, improvement of the assessments of the likely effects of military operations, automation of target identification, tracking, selection and engagement and the use of force in self-defence. In other reports, benefits such as the improvement of precision and accuracy of systems[184] as well as tasks such as mine clearance, rescue operations and the protection of civilians have been highlighted.[185]

One may wonder the result of such position regarding cyber AWS. It is interesting to note that most of the benefits raised about LAWS do not rely on physical space. Indeed a cyber AWS could also have self-destruct, self-deactivation or self-neutralization mechanisms, analyse data, conduct likely effects assessment or autonomously identify and engage a target. Of course, benefits requiring physical means such as mine clearance or civilian rescue does not apply, except following an analogy in the cyberspace. Conversely, cyber AWS would provide hereabove benefits of autonomy plus those solely related to cyberspace. Interestingly, this position emphasizes the fact that these technologies are not necessarily bad and dangerous, but can bring benefits. However, most of the debate focuses on the risk of IHL violations.

Cyber AWS, human control and the respect of IHL

The more systems become autonomous, the more fear of loss of human control increases. Human control and human-machine interaction have been key concerns within the CCW GGE debates. Even though there is a consensus on the necessity to ensure human control regarding predictability and the need of human in some decisions for IHL compliance,[186] States recognized there is still a need to determine the type and extent of human involvement or control.[187] As stated in the guiding principle (c), human-machine interactions are found at various stages of the weapon life cycle. States recognized that human control is not limited to targeting but cover a larger spectrum from research and development to post-use assessment.[188] Nevertheless, in this section, the focus will be specifically on the human control relating to the use and the targeting process of AWS regarding IHL rules.

A subsequent debate of the one opposing technological and humanist approaches are the nature of human control: direct or distributed. Some States argue there should be a direct human control where human is involved in the "concrete decisions related to the 'when and where' of the use of force need to be taken,"[189] the maintenance of communication links with the chain of command,[190] or even direct control and supervision of humans at all times and all steps.[191] One may understand the reasons of such a position that ensure the constant presence of humans in the operation. However, this may be

impracticable, especially for cyber AWS. Cyber means are rarely under the direct supervision of humans at all stages of the operation and at the targeting moment. For instance, Stuxnet worm acted without direct human control, in particular when it hit the nuclear facilities. Even concerning physical AWS, direct human control has been criticized. The United Kingdom stated that "solely relying on an operator making decisions in the heat of the moment as a panacea for human control is never the safest approach."[192] For these States, human control as a distributed control is more relevant. In this configuration, control is shared among different actors (commander, information analyst, pilot, etc.).[193] Following the idea that direct control is not necessary, some argue that more autonomy could be granted to systems if humans are at least present at one moment of the operation (especially at the programming stage) and IHL compliance is assessed.[194]

The type of human control requirement that will be adopted by CCW parties is therefore important because AWS will not act in the same manner depending on the control chosen. The issue of speed illustrates this. In some circumstances, States agreed that a system may be used autonomously if humans are too slow and the response is urgent.[195] The United States indicated that several existent defensive systems are already being used with much greater speed and accuracy than a human could achieve manually.[196] Therefore, time constraints may preclude direct human intervention and decision, and impose to rely on AWS.[197] This position is particularly welcome with respect to cyber AWS. Indeed, cyber operations take place in cyberspace with fewer constraints than in the physical world. Hence, some cyber operations occur at higher speed than a human can observe.[198] If States reject this position and require a cyber AWS operation at a human speed in order to ensure human supervision, no benefit would remain. A question that is left open is whether this potential acceptance for speed is only for defensive systems or whether it will extend to offensive means. In cyberspace, some argue that facing cyber autonomous defensive means, States will be more inclined to use such systems in offense.[199]

Whether offensive or defensive AWS, States suggested, in accordance with principle (c), several pre-programmed operational constraints as a means to ensure human-control and IHL compliance, especially principle of precaution.[200] The suggested constraints are notably related to the selected tasks, controls on the environment (target profiles, time frame, movements in an area, operating environment), parameters of use (deactivation, fail-safe mechanisms), means of interaction (overriding, abortion of the task, means of communication, monitoring and recording information mechanisms, limits on self-learning capacities).[201] However, not all States agreed on each of these constraints, especially on those requiring supervision and communication at all times. Given these constraints, States concluded that there is no "one-fit-all" human control requirement and that human-machine interactions should be established on each single use of AWS.[202] Many of these constraints are particularly suitable to cyber operations except those related to

physical environment like scope of movements.[203] For instance, the design of cyberoperation often requires to identify target profile. Moreover, cyberoperation are often unique and designed towards a precise goal.[204] Such unique design corresponds with the necessity of an ad-hoc human-machine assessment. Nonetheless, requirements such as the time frame, operating environment, overriding capabilities, permanence of means of communications may be difficult to design in cyber AWS without reducing their efficiency. Thus, if States agree on these requirements, this would reduce the possibility of potential lawful cyber AWS.

From this, it is interesting to note the relation between cyber AWS and IHL rules discussions within the CCW GGE. Because the CCW GGE does not address directly cyber AWS, few information can be extracted from the forum on this subject. So far, most States approach the question of AWS and IHL from the perspective of physical space. Therefore, it is the difference of nature between physical and cyber space that will induce adaptations. For example, the cyber operational environment may be less chaotic, at least more deterministic, than the real world and therefore cyber AWS may act with more predictability.[205] During the target process, the cyber AWS might face less difficulties due to the fact it relies on different sensors with less uncertainties, for example image recognition.[206] Similarly, concerns related to physical situations of individuals such as denial of quarter[207] or the incapacity to recognize the surrender of a combatant[208] do not apply to cyber AWS. On the other hand, the cyber space implies new elements to take into account when addressing AWS. If the environment is more deterministic and predictable, an error can be more decisive: one weakness might compromise the whole system or operation. Also, combatants are naturally out of the cyber scope but new objects are of concerns such as every connected infrastructure, especially dual use ones, or data.[209] In the same vein, an author pointed out the difficulty to suspend the protection due to civilian who participate to hostility by using cyber AWS.[210] Thus, it has been mentioned that adapted methods have to be developed for cyber AWS. For example, one author specified that to comply with principle of precaution, specific methods of reconnaissance, coordination and patching are needed to suit the cyber dimension.[211] Also, the fact that Stuxnet had evaded from to other computers are of concern regarding the geographical scope of IHL.[212]

These few elements justify the need for further discussions on cyber AWS and compliance with IHL. To date, they are not discussed by States either within the UN cyber forum or the CCW GGE due to the implicit refusal to recognize the potential threats to arising from ACC.

It is a matter of fact that autonomy will be used more and more in cyber warfare. The increased development of computing technologies, including AI, is addressing the need for faster, stealthier, innovative, and destabilizing cyber capabilities. Such methods will change the way cyber conflicts are conducted. However, improving cyber capabilities cannot be achieved in an unbridled manner. A lack of human control over increasingly ACC cannot

be done outside of any regulatory framework, especially the legal one. Therefore, the present chapter demonstrated in an innovative way that, in a sense, ACC could be characterized as LAWS. The purpose of such demonstration was to find a forum where States could discuss the future of cyber conflict and the autonomous dimension of cyber capabilities in warfare. So far, the CCW GGE is the only forum that addresses autonomy and conflict. The simplest way to provide a regulatory framework for CCA was therefore to apply the CCW GGE discussions by characterizing them as LAWS. The vast range of States positions on LAWS definition and related concepts allow to consider such assertion. From this, the analysis of ACC as cyber LAWS results in several consequences for States' positions regarding cyber means. In sum, the recognition of ACC as LAWS stimulates the debates on some current cyber issues like weapons characterization, legal review, applicability and application of IHL and human control. Furthermore, this particular exercise leads to consider the specific consequences for international law, especially IHL, regarding interactions between cyber and LAWS. While it is unlikely that Stuxnet will result in a Skynet system that endangers humanity, the use of ACC in warfare is a concern for future cyberconflicts and should be address by States whether in the UN cyber forum, the CCW GGE one or a new international forum. This thought-provoking chapter is intended to draw States' attention to this issue. The subject has been put on the table. *It will be back*.

Notes

1 Choe Sang-Hun, "North Korean leader stresses need for strong military," *The New York Times*, 15 April 2012.
2 Nicolas Falliere, Liam O Murchu, and Eric Chien, "W32.Stuxnet Dossier version 1.4," *Symentec*, 11 February 2011.
3 Rain Liivoja, Maarja Naagel, and Ann Väljataga, "Autonomous Cyber Capabilities under International Law" (Tallinn: CCDCOE, 2019), 13.
4 David Kushner, "The real story of Stuxnet," *IEEE Spectrum*, 26 February 2013.
5 Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: W.W. Norton & Company, 2018), 258–260.
6 It is interesting to note that Stuxnet is generally not considered as an AI system. Of course, this depends on the definition of AI adopted. A broad definition of AI, including any computer algorithm, would refer to Stuxnet. A narrow definition referring to specific algorithms such as machine learning algorithms would exclude Stuxnet from the definition of an AI system.
7 United Nations General Assembly (UNGA), "Developments in the field of information and telecommunications in the context of international security," *United Nations, A/RES/53/70*, 4 January 1999.
8 For more information on the evolution of the number of members of different Groups of Governmental Experts (GGE), see UN Office for Disarmament Affairs (UNODA), "Developments in the field of information and telecommunications in the context of international security," Fact Sheet, (United Nations, July 2019).
9 UNGA, "Report of the group of governmental experts on developments in the field of information and telecommunications in the context of international security," U*nited Nations, A/65/201*, 30 July 2010.

10 UNGA, "Report of the group of governmental experts on developments in the field of information and telecommunications in the context of international security," *United Nations, A/68/98*, 24 June 2013.
11 UNGA, "Report of the group of governmental experts on developments in the field of information and telecommunications in the context of international security," *United Nations, A/70/174*, 22 July 2015.
12 UNGA, "Report of the group of governmental experts on developments in the field of information and telecommunications in the context of international security," *United Nations, A/76/135*, 14 July 2021.
13 UNGA, "Final substantive report of the open-ended working group [OEWG] on developments in the field of information and telecommunications in the context of international security," *United Nations, A/AC.290/2021/CRP.2*, 10 March 2021.
14 For more information on the Group of Governmental Experts (GGE) on Emerging Technologies in the Area of lethal autonomous weapon systems (LAWS), see the UN website: https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/.
15 For all reports, working papers and statements, see the UNODA Meetings Place: https://meetings.unoda.org/meeting/ccw-gge-2020/.
16 Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons (CCW), GGE LAWS (CCW GGE LAWS), "Report of the 2019 session of the group of governmental experts on emerging technologies in the area of lethal autonomous weapons systems," *CCW/GGE.1/2019/3*, 25 September 2019, Annex IV, "Guiding principles."
17 Alessandro Guarino, "Autonomous intelligent agents in cyber offence," in *5th International Conference on Cyber Conflict (CYCON 2013)* (Tallinn, Estonia: IEEE, 2013), 5.
18 Defense Science Board, "Task force report: Resilient military systems and the advanced cyber threat" (United States Department of Defense, January 2013), 19.
19 CCW GGE LAWS, "Report […] CCW/GGE.1/2019/3," para. (f).
20 Caitríona Heinl, "Maturing autonomous cyber weapons systems: Implications for international security cyber and autonomous weapons systems regimes," in *Oxford Handbook of Cyber Security* (Oxford University Press, 2018), 3.
21 Caitríona Heinl, "Artificial (intelligent) agents and active cyber defence: policy implications," in *2014 6th International Conference On Cyber Conflict (CyCon 2014)* (Tallinn: NATO CCDCOE Publications, 2014); Guarino, "Autonomous intelligent agents."
22 Tim McFarland, "The concept of autonomy," in *Autonomous Cyber Capabilities under International Law* (NATO CCDCOE Publications, 2021), 17.
23 Ibid., 22.
24 Tanel Tammet, "Autonomous cyber defence capabilities," in *Autonomous Cyber Capabilities under International Law* (NATO CCDCOE Publications, 2021), 42.
25 Ibid., 36.
26 Ibid., 36–37.
27 Ibid., 37.
28 Guarino, "Autonomous intelligent agents," 8.
29 Tammet, "Autonomous cyber defence capabilities," 39–42.
30 McFarland, "The concept of autonomy," 3.
31 Liivoja, Naagel and Väljataga, "Autonomous cyber capabilities," 10.
32 McFarland, "The concept of autonomy," 23–24.
33 Ibid., 24–25.
34 Tammet, "Autonomous cyber defence capabilities," 39; Guarino, "Autonomous intelligent agents," 2.
35 Jia Song and Jim Alves-Foss, "The DARPA cyber grand challenge: A competitor's perspective," *IEEE Security & Privacy* 13, no. 6 (2015): 72–76.

36 Guarino, "Autonomous intelligent agents," 6–8.
37 Heinl, "Maturing autonomous cyber weapons systems," 4; UNIDIR, "The weaponization of increasingly autonomous technologies: Autonomous weapon systems and cyber operations," 2017, 1–5; UNIDIR, "Side event on cyber and autonomous weapons," 14 October 2015.
38 UNGA, "Chair's summary of the open-ended working group on developments in the field of information and telecommunications in the context of international security," *United Nations, A/AC.290/2021/CRP.3*, 10 March 2021, §7.
39 United States Department of Defense, "Autonomy in weapons systems," *Department of Defense directive 3000.09*, 21 November 2012.
40 Ibid., §2.b.
41 Scharre, *Army of None*, 326: "This wasn't because we thought autonomous cyberweapons were uninteresting or unimportant when we wrote the directive. It was because we knew bureaucratically it would be hard enough simply to create a new policy on autonomy. Adding cyber operations would have multiplied the complexity of the problem, making it very likely we would have accomplished nothing at all."
42 CCW GGE LAWS, "Presentation at the United Nations Convention on certain conventional weapons by Paul Scharre," *Expert Statement*, 13 April 2015: 4. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2015)/Scharre%2Bpresentation%2Btext.pdf.
43 CCW GGE LAWS, "The right to life and the Martens clause by Patrick Lin," *Expert Statement*, 5 April 2015: 6. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2015)/24%2BPatrick%2BLin_Patrick%2BSS.pdf.
44 CCW GGE LAWS, "Statement of the UN institute for disarmament research by Kerstin Vignard," *Expert Statement*, 16 April 2016: 2. https://www.unidir.org/files/medias/pdfs/unidir-s-statement-to-the-ccw-informal-meeting-of-experts-on-lethal-autonomous-weapon-systems-eng-0–648.pdf.
45 CCW GGE LAWS, "Chairperson's summary," Working paper, *CCW/GGE.1/2020/WP.7*, 19 April 2021, Annex III: 74 (Portugal §31).
46 CCW GGE LAWS, "Operationalizing the guiding principles: A roadmap for the GGE on LAWS," Working paper submitted by Brazil, CCW/GGE.1/2020/WP.3, 6 August 2020: §4–5. https://documents.unoda.org/wp-content/uploads/2020/08/CCW-GGE.1–2020-WP.3-.pdf; "Working paper by the Bolivarian Republic of Venezuela on behalf of the Non-Aligned Movement (NAM) and other states parties to the Convention on Certain Conventional Weapons (CCW)," CCW/GGE.1/2020/WP.5, 14 September 2020: §20. http://undocs.org/CCW/GGE.1/2020/WP.5; "Potential opportunities and limitations of military uses of lethal autonomous weapons systems," Working paper submitted by Russia, CCW/GGE.1/2019/WP.1, 15 March 2019: §5. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/CCW.GGE.1.2019.WP.1_R%2BE.pdf; "Statement by the Netherlands," 25 April 2019. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/5d%2BNL%2Bstatement%2BCharacterization-final.pdf; "Statement by Greece," March 2019. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/GGE%2BLAWS%2BSTATEMENT%2BBY%2B%2BGREECE-Charecteristics%2Bof%2BLAWS.pdf; "Position Paper," Working paper submitted by China, CCW/GGE.1/2018/WP.7, 11 April 2018: 1. https://docs-library.unoda.org/

Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/CCW_GGE.1_2018_WP.7.pdf; "Statement Estonia," April 2017. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_LAWS6a_Estonia.pdf; "Statement of Switzerland," 11 April 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_LAWS6b_Switzerland.pdf; "Statement by the United Kingdom," 10 April 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_LAWS6a_UK.pdf; "Statement by Pakistan," 27 August 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_GGE%2BLAWS%2B2_6a_Pakistan.pdf; "Statement by Ireland," August 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_GGE%2BLAWS%2B2_6a_Ireland.pdf; "Characteristics of Lethal Autonomous Weapons Systems," Working paper submitted by the United States, CCW/GGE.1/2017/WP.7, 10 November 2017: §6. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2017)/2017_GGEonLAWS_WP7_USA.pdf; "Japan's views on issues relating to LAWS," Working paper submitted by Japan, 2016. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2016)/2016_LAWS%2BMX_CountryPaper%2BJapan.pdf; "Elements supporting the prohibition of lethal autonomous weapons systems," Working paper submitted by the Holy See, 7 April 2016. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2016)/2016_LAWSMX_CountryPaper_Holy%2BSee.pdf; "Non-paper characterization of LAWS," Working paper submitted by France, April 2016. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2016)/2016_LAWSMX_CountryPaper_France%2BCharacterizationofaLAWS.pdf; "Statement by Italy," April 2016. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2016)/2016_LAWS_MX_towardsaworkingdefinition_statements_Italy.pdf.

47 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex II: §18; "Statement by India," 25 March 2019. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/25%2BMarch%2B2019%2B-%2B5%28d%29.pdf; "Position Paper […] CCW/GGE.1/2018/WP.7," 1 (see n46).

48 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 36 (Cuba §14); "Report of the 2018 session of the group of governmental experts on emerging technologies in the area of lethal autonomous weapons systems," *CCW/GGE.1/2018/3*, 23 October 2018: §22(a).

49 CCW GGE LAWS, "Examination of various dimensions of emerging technologies in the area of lethal autonomous weapons systems, in the context of the objectives and purposes of the convention," Working paper submitted by the Netherlands, CCW/GGE.1/2017/WP.2, 9 October 2017. https://undocs.org/ccw/gge.1/2017/WP.2; "Japan's views on issues," §5 (see n46).

50 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3."

51 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," 19 (Australia); "Statement by Germany," 25 March 2019: §4. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/20190325%2BStatement

2%2BGermany%2BGGE%2BLAWS.pdf; "Statement by the Netherlands," 2018: 1. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_LAWS6a_Netherlands.pdf.

52 CCW GGE LAWS, "Statement by Pakistan" (see n46).

53 CCW GGE LAWS, "Characteristics of Lethal […] CCW/GGE.1/2017/WP.7," §3 (see n46).

54 CCW GGE LAWS, "Statement by Pakistan" (see n46); "Statement by Switzerland," 10 April 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_LAWS6a_Switzerland.pdf.

55 CCW GGE LAWS, "Statement by Switzerland," 27 August 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018.08.27%2BGGE%2BLAWS_Switzerland_Item%2B6a.pdf.

56 CCW GGE LAWS, "Statement of the European Union," 25–26 March 2019: 1–2. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/ALIGNED%2B-%2BLAWS%2BGGE%2BEU%2Bstatement%2BHuman%2BElement.pdf.

57 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex II: §19, Annex III: 46 (Germany §2); "Examination of various dimensions […] CCW/GGE.1/2017/WP.2," §2 (see n49); "Characteristics of Lethal […] CCW/GGE.1/2017/WP.7," §6 (see n46).

58 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex II: §20, Annex III: 76 (Russia), 94–102 (United States), 106 (Non-Aligned Movement); "Russia's Approaches to the Elaboration of a Working Definition and Basic Functions of Lethal Autonomous Weapons Systems in the Context of the Purposes and Objectives of the Convention," Working paper submitted by Russia, CCW/GGE.1/2018/WP.6, 4 April 2018: §3. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/CCW_GGE.1_2018_WP.6_E.pdf; "Statement by Italy," April 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_LAWS6a_Italy.pdf; "Position Paper […] CCW/GGE.1/2018/WP.7" (see n46); "Statement by Bulgaria," April 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_LAWS6a_Bulgaria.pdf; "Characteristics of Lethal […] CCW/GGE.1/2017/WP.7," §6 (see n46).

59 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," 94–102 (United States).

60 "[T]he Chairperson will convene in 2014 a four-day informal Meeting of Experts, from 13 to 16 May 2014, to discuss the questions related to emerging technologies in the area of lethal autonomous weapons systems, in the context of the objectives and purposes of the Convention." CCW, Meeting of the High Contracting Parties, "Final report," CCW/MSP/2013/10, 16 December 2013: §32. https://undocs.org/CCW/MSP/2013/10.

61 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," 79.

62 Ibid., 46–50 (Germany), 76–79 (Russia).

63 Ibid., 46–50 (Germany), 70–76, 76–79 (Russia).

64 Ibid., CCW/GGE.1/2020/WP.7," 94–102 (United States).

65 CCW, art. 1.

66 CCW GGE LAWS, "Weapons Review Mechanisms," Working paper submitted by the Netherlands and Switzerland, CCW/GGE.1/2017/WP.5, 7 November 2017: §27. https://undocs.org/ccw/gge.1/2017/WP.5.
67 Ibid.
68 CCW GGE LAWS, "Categorizing lethal autonomous weapons systems – A technical and legal perspective to understanding LAWS," Working paper submitted by Estonia and Finland, CCW/GGE.2/2018/WP.2, 24 August 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_GGE%2BLAWS_August_Working%2BPaper_Estonia%2Band%2BFinland.pdf.
69 CCW GGE LAWS, "Statement by Estonia," 25–29 March 2019. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/LAWS%2BGGE%2B2019%2BI%2B-%2BEstonia%2B-%2BAgenda%2Bitem%2B5%28b%29.pdf.
70 Michael N. Schmitt and Liis Vihul (eds). *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge: Cambridge University Press, 2017), 452.
71 Ibid., 415.
72 UNGA, "Réponse de la France à la résolution 73/27 relative aux « Progrès de l'informatique et des télécommunications et sécurité internationale » et à la résolution 73/266 relative à « Favoriser le comportement responsable des États dans le cyberespace dans le contexte de la sécurité internationale," 2019. https://unoda-web.s3.amazonaws.com/wp-content/uploads/2019/09/France-2019.pdf.
73 See: Egypt, UNGA OEWG, "Compendium of statements in explanation of position on the final report," A/AC.290/2021/INF/2, 20 March 2021: 34–35. https://front.un-arm.org/wp-content/uploads/2021/04/A-AC.290–2021-INF-2.pdf; Non-Aligned Movement, UNGA OEWG, "NAM Statement," February 2021. https://front.un-arm.org/wp-content/uploads/2021/02/NAM-Statement-Informal-Consultation-OEWG-on-ICT.pdf; Zimbabwe, UNGA OEWG, "Zimbabwe Statement," 22 February 2021. https://front.un-arm.org/wp-content/uploads/2021/02/ZIMBABWE-GENERAL-STATEMENT-THE-INFORMAL-CONSULTATIONS-OF-THE-OEWG-22-FEBRUARY-2021.pdf; Iran, UNGA OEWG, "Iran Statement," 2019. https://unoda-web.s3.amazonaws.com/wp-content/uploads/2019/09/iran-submission-oewg-sep-2019.pdf; Pakistan, UNGA OEWG, "Pakistan Statement," March 2020. https://unoda-web.s3.amazonaws.com/wp-content/uploads/2020/03/working-paper-pakistan.pdf.
74 The Netherlands, UNGA OEWG, "Appendix: International Law in cyberspace," February 2020. https://unoda-web.s3.amazonaws.com/wp-content/uploads/2020/02/appendix-Internaional-law-in-cyberspace-kingdom-of-the-netherlands.pdf; Brazil, UNGA OEWG, "Brazil Statement," 8 April 2020. https://front.un-arm.org/wp-content/uploads/2020/04/comments-by-brazil-on-the-pre-draft-report-of-cyber-oewg-8-apr-2020.pdf.
75 UNGA OEWG, "United Kingdom Statement," February 2020. https://unoda-web.s3.amazonaws.com/wp-content/uploads/2020/03/20200303-uk-national-contribution-oewg2.pdf; UNGA OEWG, "Cuba Statement," April 2020. https://front.un-arm.org/wp-content/uploads/2020/04/considerations-on-the-initial-pre-draft-of-the-oewg-cybersecurity-cuba-15-april.pdf.
76 UNGA OEWG, "Russia Statement," April 2020. https://front.un-arm.org/wp-content/uploads/2020/04/russian-commentary-on-oweg-zero-draft-report-eng.pdf.

77 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex II: §20.
78 Ibid., Annex III: 36 (Cuba §12).
79 CCW GGE LAWS, "Possible outcome of 2019 group of governmental experts and future actions of international community on lethal autonomous weapons systems," Working paper submitted by Japan, CCW/GGE.1/2019/WP.3, 22 March 2019: §13. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2017)/2017_GGEonLAWS_WP9_Switzerland.pdf.
80 Ibid., §12.
81 CCW GGE LAWS, "Position Paper […] CCW/GGE.1/2018/WP.7" (see n46).
82 CCW GGE LAWS, "Statement of Switzerland (11 April 2018)" (see n46).
83 CCW GGE LAWS, "Statement by Ireland," 25 March 2019. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/Statement%2Bby%2BIreland%2Bunder%2BAgenda%2BItem%2B5b%2B-%2BCharactertisation%2B-%2BGGE%2BMarch%2B2019%2B-%2BFinal.pdf.
84 CCW GGE LAWS, "Categorizing lethal autonomous […] CCW/GGE.2/2018/WP.2" (see n68).
85 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 28 (Austria §11).
86 CCW GGE LAWS, "Statement by Ireland" (see n83).
87 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3," Annex III: §4.
88 CCW GGE LAWS, "A "compliance-based" approach to autonomous weapon systems," Working paper submitted by Switzerland, CCW/GGE.1/2017/WP.7, 10 November 2017: §27. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2017)/2017_GGEonLAWS_WP9_Switzerland.pdf.
89 Article 49, Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, June 8, 1977, 1125 U.N.T.S. 3. (hereinafter AP I).
90 CCW GGE LAWS, "A "compliance-based" approach […] CCW/GGE.1/2017/WP.7," §27 (see n88).
91 Schmitt and Vihul, *Tallinn Manual 2.0*, 415.
92 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," §2; "Statement by Estonia (2017)" (see n46).
93 CCW GGE LAWS, "Statement by Ireland" (see n83).
94 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 19 (Australia §5); "Statement by the United Kingdom" (see n46); "Statement by Estonia (2017)" (see n46); "Statement by Italy" (see n58); "Statement by Norway," November 2017. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2017)/2017_GGE%2BLAWS_Statement_Norway.pdf; "Towards a definition of lethal autonomous weapons systems," Working paper submitted by Belgium, CCW/GGE.1/2017/WP.3, 7 November 2017: §4–6. https://undocs.org/ccw/gge.1/2017/WP.3; "For consideration by the Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS)," Working paper submitted by France and Germany, CCW/GGE.1.2017/WP.4, 7 November 2017: §6. https://undocs.org/ccw/gge.1/2017/WP.4.
95 CCW GGE LAWS, "Position Paper […] CCW/GGE.1/2018/WP.7" (see n46); "For consideration by […] CCW/GGE.1/2017/WP.4," §6 (see n94); "Statement by Bulgaria" (see n58).

96 CCW GGE LAWS, "Statement by the United Kingdom" (see n46); "Statement by Italy" (see n58); "Examination of various dimensions […] CCW/GGE.1/2017/WP.2," §2 (see n49); "Japan's views on issues," §3 (see n46).
97 CCW GGE LAWS, "Statement by Russia," 22 November 2017. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2017)/2017_GGE%2BLAWS_Statement_Russia.pdf; CCW, Informal Meeting of Experts LAWS, "Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)," 2016: §13. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2016)/Report-LAWS_2016_AdvancedVersion.pdf.
98 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 37 (Cuba §16–17).
99 CCW GGE Laws, "Potential opportunities […] CCW/GGE.1/2019/WP.1," §2–5 (see n46).
100 See also, CCW GGE LAWS, "Statement by Norway" (see n94).
101 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," §20; "Statement Estonia (2017)" (see n46); "Categorizing lethal autonomous […] CCW/GGE.2/2018/WP.2" (see n68).
102 CCW GGE LAWS, "Statement by Brazil," 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_LAWS6a_Brazil1.pdf.
103 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3," Annex III: §4.
104 CCW GGE LAWS, "Statement by United Kingdom," 10 April 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_LAWS6a_UK.pdf.
105 CCW GGE LAWS, "Statement of Switzerland (11 April 2018)" (see n46).
106 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3," Annex III: §6.
107 CCW GGE LAWS, "Statement of Switzerland (April 2018)" (see n46).
108 CCW GGE LAWS, "Statement by Switzerland (August 2018)" (see n55); "Categorizing lethal autonomous […] CCW/GGE.2/2018/WP.2" (see n68).
109 CCW GGE LAWS, "Statement Estonia," August 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_GGE%2BLAWS%2B2_6a_Estonia.pdf.
110 CCW GGE LAWS, "Statement of Switzerland (11 April 2018)" (see n46); "Statement Estonia (2018)" (see n109); "Statement by Ireland" (see n83); "Statement by Greece" (see n46).
111 CCW GGE LAWS, "Statement of Switzerland (April 2018)" (see n46); "Statement Estonia (2018)" (see n46); "Statement by the United Kingdom" (see n46); "Statement by Switzerland (August 2018)" (see n55).
112 CCW GGE LAWS, "Statement by the United Kingdom" (see n46).
113 CCW GGE LAWS, "Statement by the International Committee of the Red Cross," April 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/CCW%2BGGE%2B-%2BApril%2B2018%2B-%2BICRC%2Bstatement%2B-%2Bcharacterisation%2B6a.pdf.
114 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3," Annex III: §47.
115 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3," Annex III: §4; "Statement by Italy" (see n58).
116 CCW GGE LAWS, "Statement of Switzerland (11 April 2018)" (see n46).

117 CCW GGE LAWS, "Statement by Estonia (2017)" (see n46).
118 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," §19; "Report of the 2018 session […] CCW/GGE.1/2018/3," §22; "Statement by the Netherlands (2019)" (see n46); "Statement by the United Kingdom" (see n46); "Statement by Austria," April 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_LAWS6a_Austria.pdf. On the technical evolution: "Statement by the United Kingdom," March 2019. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/20190318-5%28b%29_Characterisation_Statement.pdf; "Statement by the Holy See," August 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_GGE%2BLAWS%2B2_6a_Holy%2BSee.pdf. On the legal obligation: CCW GGE LAWS, "Statement by the United Kingdom" (see n118); "Towards a definition […] CCW/GGE.1/2017/WP.3," §4–5 (see n94).
119 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3," Annex III: §4, §22; "Statement by Austria" (see n118).
120 McFarland, "The Concept of Autonomy," 14.
121 It should be recalled here that the different conceptions of autonomy, and related concepts to, should not be seen as opposed. One State may consider that LAWS are only fully autonomous weapons, i.e., weapons with autonomy in their critical functions that work without human intervention once activated while another State may views LAWS as an automated system with autonomy in its targeting function without human supervision. A myriad of definitions is possible given all the criteria and elements mentioned by States.
122 Scharre, *Army of None*, 326; Liivoja, Naagel and Väljataga, "Autonomous Cyber Capabilities," 12.
123 Monica Kaminska, Dennis Broeders, and Fabio Cristiano, "Limiting viral spread: Automated cyber operations and the principles of distinction and discrimination in the grey zone," in *Going Viral (13th International Conference on Cyber Conflict)*, eds. T. Jančárková, L. Lindström, G. Visky, and P. Zotz (Tallinn: NATO CCDCOE Publication, 2021), 63; Tammet, "Autonomous cyber defence capabilities," 38.
124 McFarland, "The concept of autonomy," 129.
125 Marco De Falco, "Stuxnet Facts Report: A Technical and Strategic Analysis" (NATO CCDCOE, 2012), 21.
126 CCW GGE LAWS, "Report […] CCW/GGE.1/2019/3," §(a): "International humanitarian law continues to apply fully to all weapons systems, including the potential development and use of lethal autonomous weapons systems".
127 CCW, art.1.
128 NATO, "Warsaw Summit Communiqué," 9 July 2016. https://www.nato.int/cps/en/natohq/official_texts_133169.htm.
129 UNGA, "Statement by the EU," 26 October 2018. https://eeas.europa.eu/delegations/un-new-york/52894/eu-statement-%E2%80%93-united-nations-1st-committee-thematic-discussion-other-disarmament-measures-and_en.
130 In 2021, within the UN GGE on cyber, numerous States shared their views on how international law applies to cyber and recognized the applicability of IHL (notably Australia, Brazil, Estonia, Germany, Japan, Kenya, Netherlands, Norway, Romania, Singapore, Switzerland, United Kingdom and United States), see UNGA, "Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States'," 13 July 2021, A/76/136: 6, 23, 29, 39, 54, 59, 65, 79, 93, 97, 132, 148. See also, for example: France, "International

Law Applied to Operations in Cyberspace" (Ministère des Armées, October 2019), 21; United States, "Law of War Manual" (Department of Defense, 2016), 985–1000.

131 Schmitt and Vihul, *Tallinn Manual 2.0*, 375.
132 UNGA, "Report […] A/70/174," §28(d).
133 It is noteworthy that the report does not refer to the precautionary principle.
134 UNGA GGE Cyber, "Statement by Cuba," 23 June 2017. https://www.justsecurity.org/wp-content/uploads/2017/06/Cuban-Expert-Declaration.pdf.
135 CCW, Informal Meeting of Experts LAWS, "Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)," CCW/MSP/2015/3, 2 June 2015: §52. https://undocs.org/en/CCW/MSP/2015/3.
136 Michael N. Schmitt and Liis Vihul, "International Cyber Law Politicized: The UN GGE's Failure to Advance Cyber Norms," *Just Security* (blog), 30 June 2017.
137 UNGA, "Chair's Summary […] A/AC.290/2021/CRP.3," §12.
138 UNGA, "Advance copy, report of the group of governmental experts on advancing responsible state behaviour in cyberspace in the context of international security," 28 May 2021: §71(f). https://front.un-arm.org/wp-content/uploads/2021/06/final-report-2019–2021-gge-1-advance-copy.pdf.
139 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 86 (Switzerland §7).
140 Ibid., Annex III: 46 (Germany §3), 70 (Portugal, §10).
141 UNGA, "Official compendium of […] A/76/136" (see n130).
142 AP I, art. 36.
143 In this regard, it is interesting to note that the CCW discussions put this point on the table. The United States refused to recognize this article as part of customary law whereas Russia did. CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 76 (Russia §11), 94 (United States §13).
144 Schmitt and Vihul, *Tallinn Manual 2.0*, 465.
145 CCW GGE LAWS, "Report […] CCW/GGE.1/2019/3," §(e): "In accordance with States' obligations under international law, in the study, development, acquisition, or adoption of a new weapon, means or method of warfare, determination must be made whether its employment would, in some or all circumstances, be prohibited by international law".
146 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," §11.
147 Ibid., §10, Annex II: §13, Annex III: 28 (Austria §16), 36 (Cuba §23); CCW GGE LAWS, "Strengthening of the review mechanism of a new weapon, means and method of warfare," Working paper by Argentina, CCW/GGE.1/2018/WP.2, 4 April 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/CCW_GGE.1_2018_WP.2.En.pdf; "Weapons Review Mechanisms […] CCW/GGE.1/2017/WP.5" (see n66).
148 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," §15(1), Annex III: 60 (Netherlands §12), 91 (United Kingdom §7–8), 94 (United State §15); "Questionnaire on the Legal Review Mechanisms of New Weapons, Means and Methods of Warfare," Working paper submitted by Argentina, CCW/GGE.1/2019/WP.6, 29 March 2019: §3; "Report of the 2018 session […] CCW/GGE.1/2018/3," Annex III: §9; "Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems," CCW/GGE.1/2017/3, 22 December 2017: Annex, II, §23. https://undocs.org/CCW/GGE.1/2017/3; "For consideration by […] CCW/GGE.1/2017/WP.4," §13–22 (see n94).

149 CCW GGE LAWS, "Questionnaire on the Legal […] CCW/GGE.1/2019/WP.6," §3 (see n148); "Strengthening of the review […] CCW/GGE.1/2018/WP.2," §10–11; AP I, art 84 (see n147).
150 On confidence-building measure: CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," §10–11, Annex II: §14. On trust: CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 54 (Italy §1(b)). On transparence: CCW GGE LAWS, "Statement of the European Union," 1–2 (see n56); "Report of the 2018 session […] CCW/GGE.1/2018/3," Annex III: §44.
151 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," §10–11, Annex II: §14.
152 Ibid.; CCW GGE LAWS, "Strengthening of the review […] CCW/GGE.1/2018/WP.2," §10–11 (see n147).
153 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3," Annex III: §44; "Questionnaire on the Legal […] CCW/GGE.1/2019/WP.6," §3 (see n148).
154 CCW GGE LAWS, "For consideration by […] CCW/GGE.1/2017/WP.4," §9(b) (see n94).
155 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," §10, Annex II: §14, Annex III: 76 (Russia §11); "Strengthening of the review […] CCW/GGE.1/2018/WP.2," §10–11 (see n147).
156 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 54 (Italy §1), 60 (Netherlands §12), 94 (United States §15); "Report […] CCW/GGE.1/2019/3," §17(i).
157 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 2 (Austria).
158 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3," Annex III: §39.
159 CCW GGE LAWS, "The Australian Article 36 Review Process," CCW/GGE.2/2018/WP.6, 30 August 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_GGE%2BLAWS_August_Working%2Bpaper_Australia.pdf.
160 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 46 (Germany §15).
161 CCW GGE LAWS, "Statement of the Netherlands," 26 April 2019: 3. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/5a%2BNL%2BStatement%2BLegal%2BChallenges-final.pdf.
162 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 76 (Russia §12); "Potential opportunities and limitations of military uses of lethal autonomous weapons systems," Working paper submitted by Russia, CCW/GGE.1/2017/WP.9, 10 November 2017: §9. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2017)/2017_GGEonLAWS_WP9_Switzerland.pdf.
163 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 83 (Sweden §10–12).
164 Vincent Boulanin and Maaike Verbruggen, "Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies" (SIPRI, 2017).
165 UNGA, "Final Substantive Report [OEWG]" A/AC.290/2021/CRP.2, §41–47; UNGA, "Advance copy," §74–86 (see n138).
166 Schmitt and Vihul, *Tallinn Manual 2.0*, 365.
167 UNIDIR, "The Weaponization of," 1–3.

168 UNGA GGE Cyber, "Australia's submission on international law to be annexed to the report of the 2021 Group of Governmental Experts on Cyber," June 2021: 4. https://www.internationalcybertech.gov.au/sites/default/files/2021–06/Australia Annex - Final%2C as submitted to GGE Secretariat.pdf.
169 CCW GGE LAWS, "The Australian Article […] CCW/GGE.2/2018/WP.6" (see n159).
170 UNGA, "Official compendium of […] A/76/136," 23, 103 (see n130).
171 Schmitt and Vihul, *Tallinn Manual 2.0*, 464; David Wallace, "Cyber Weapon Reviews under International Humanitarian Law" (NATO CCDCOE, 2018); Gary Brown and Andrew Metcalf, "Easier Said Than Done: Legal Reviews of Cyber Weapons," *Journal of National Security Law and Policy* 7 (2014)'; Boulanin and Verbruggen, "Article 36 Reviews."
172 For a detail study, although quite technical, on that subject see Alec Tattersall and Damian Copeland, "Reviewing Autonomous Cyber Capabilities," in *Autonomous Cyber Capabilities under International Law* (NATO CCDCOE Publications, 2021).
173 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3," §26.
174 CCW, Informal Meeting of Experts LAWS, "Report of the 2015 Informal Meeting […] CCW/MSP/2015/3," §34 (see n135).
175 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3," §23(c).
176 Christopher M. Kovach, "Beyond Skynet: Reconciling Increased Autonomy in Computer-Based Weapons Systems with the Laws of War," *Air Force Law Review* 71 (2014): 245.
177 UNIDIR, "The Weaponization of," 2; Heinl, "Maturing Autonomous Cyber Weapons Systems," 16.
178 CCW GGE LAWS, "Report […] CCW/GGE.1/2019/3," § (c): "Human-machine interaction, which may take various forms and be implemented at various stages of the life cycle of a weapon, should ensure that the potential use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems is in compliance with applicable international law, in particular IHL. In determining the quality and extent of human-machine interaction, a range of factors should be considered including the operational context, and the characteristics and capabilities of the weapons system as a whole."
179 For more details on that point, see Vincent Boulanin, Laura Bruun, and Netta Goussac, "Autonomous Weapon Systems and International Humanitarian Law. Identifying Limits and the Required Type and Degree of Human–Machine Interaction" (SIPRI, 2021), 18–20.
180 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 23–28 (Austria, Belgium, Brazil, Chile, Ireland, Germany, Luxembourg, Mexico and New-Zealand), 86–91 (Switzerland); "Working paper […] CCW/GGE.1/2020/WP.5" (see n46).
181 Heinl, "Maturing Autonomous Cyber Weapons Systems," 17.
182 CCW GGE LAWS, "Report […] CCW/GGE.1/2019/3," para. (h).
183 CCW GGE LAWS, "Humanitarian benefits of emerging technologies in the area of lethal autonomous weapon systems," Working paper submitted by the United-States, CCW/GGE.1/2018/WP.4, 28 March 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/CCW_GGE.1_2018_WP.4.pdf.
184 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," §12.
185 CCW, Informal Meeting of Experts LAWS, "Report of the 2016 Informal Meeting," §35 (see n97).

186 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," §6, §11, Annex II: §8.
187 Ibid., §21.
188 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3," Annex III: §10. The Chair shared a 'sunrise slide' of touch points in the human-machine interface: (0) political direction in the pre-development phase; (1) research and development; (2) testing, evaluation and certification; (3) deployment, training, command and control; (4) use and abort; (5) post-use assessment.
189 Ibid., Annex III, §19; CCW GGE LAWS, "Statement by Greece" (see n46).
190 CCW GGE LAWS, "Report of the 2018 session […] CCW/GGE.1/2018/3," Annex III: §19.
191 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," 106 (Non-Aligned Movement §27), 68 (Poland §5).
192 CCW GGE LAWS, "United Kingdom expert paper: The human role in autonomous warfare," Working paper submitted by United Kingdom, CCW/GGE.1/2020/WP.6, 18 November 2015: §15. https://undocs.org/CCW/GGE.1/2020/WP.6.
193 CCW GGE LAWS, "Statement of the European Union," 1–2 (see n56); "Statement of the Netherlands," 2–3 (see n161); "Statement of Switzerland (11 April 2018)" (see n46).
194 This position clearly refers to the technological conception, see McFarland, "The Concept of Autonomy," 26–34.
195 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 81 (Spain §14); "Operationalizing the guiding principles […] CCW/GGE.1/2020/WP.3," §26 (see n46)
196 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 81 (Spain §14); "Humanitarian benefits […] CCW/GGE.1/2018/WP.4," §10 (see n183).
197 CCW GGE LAWS, "United Kingdom Expert paper […] CCW/GGE.1/2020/WP.6," §17 (see n192).
198 McFarland, "The concept of autonomy," 18.
199 Eric Messinger, "Is it possible to ban autonomous weapons in cyberwar?" *Just Security* (blog), 15 January 2015.
200 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," Annex III: 23 (Austria, Belgium, Brazil, Chile, Ireland, Germany, Luxembourg, Mexico, and New-Zealand §8).
201 Ibid., Annex III: 19 (Austria §14), 23 (Austria, Belgium, Brazil, Chile, Ireland, Germany, Luxembourg, Mexico, and New-Zealand §14), 34 (Costa Rica §14), 41 (Finland §7), 70 (Portugal §17), 81 (Spain §13), 83 (Sweden §7), 86 (Switzerland §6, §15); "Operationalizing the guiding principles […] CCW/GGE.1/2020/WP.3," §24 (see n46); "United Kingdom Expert paper […] CCW/GGE.1/2020/WP.6," §17 (see n192); "Statement of the Netherlands" (see n161).
202 CCW GGE LAWS, "Chairperson's summary […] CCW/GGE.1/2020/WP.7," 51 (Israel §11), 44 (France), 81 (Spain §11), Annex II: §12.
203 Although an analogy with physical systems could compare physical movements to "movements" through different servers.
204 Tattersall and Copeland, "Reviewing autonomous cyber capabilities," 225.
205 Guarino, "Autonomous intelligent agents," 6.
206 Liivoja, Naagel and Väljataga, "Autonomous cyber capabilities," 28.
207 CCW GGE LAWS, "A "compliance-based" approach […] CCW/GGE.1/2017/WP.7," §11 (see n88).
208 CCW GGE LAWS, "Advanced version. report of the 2016 informal meeting of experts on lethal autonomous weapons systems (LAWS)," 2016: §45. https://

docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_Informal_Meeting_of_Experts_(2016)/ReportLAWS_2016_AdvancedVersion.pdf.

209 Kubo Mačák, "Military objectives 2.0: The case for interpreting computer data as objects under international humanitarian law," *Israel Law Review* 48, no. 1 (March 2015).

210 Tassilo V.P. Singer, "Update to revolving door 2.0: The extension of the period for direct participation in hostilities due to autonomous cyber weapons," in *2017 9th International Conference on Cyber Conflict (CyCon)* (IEEE, 2017).

211 Peter Margulies, "A moment in time: Autonomous cyber capabilities, proportionality and precautions," in *Autonomous Cyber Capabilities under International Law*, eds. Rain Liivoja, and Ann Väljataga (NATO CCDCOE Publications, 2021), 172–179.

212 Liivoja, Naagel and Väljataga, "Autonomous cyber capabilities," 27.

Bibliography

Boulanin, Vincent, Laura Bruun, and Netta Goussac. "Autonomous weapon systems and international humanitarian law. Identifying limits and the required type and degree of human–machine interaction." *SIPRI*, 2021. https://www.sipri.org/publications/2021/other-publications/autonomous-weapon-systems-and-international-humanitarian-law-identifying-limits-and-required-type.

Boulanin, Vincent, and Maaike Verbruggen. "Article 36 reviews: Dealing with the challenges posed by emerging technologies." *SIPRI*, 2017. https://www.sipri.org/publications/2017/other-publications/article-36-reviews-dealing-challenges-posed-emerging-technologies.

Brown, Gary, and Andrew Metcalf. "Easier said than done: Legal reviews of cyber weapons." *Journal of National Security Law and Policy* 7 (2014): 115–138.

Choe, Sang-Hun. "North Korean leader stresses need for strong Military." *The New York Times*, 15 April 2012. https://www.nytimes.com/2012/04/16/world/asia/kim-jong-un-north-korean-leader-talks-of-military-superiority-in-first-public-speech.html.

Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons (CCW), GGE LAWS (CCW GGE LAWS). "Report of the 2018 session of the group of governmental experts on emerging technologies in the area of lethal autonomous weapons systems." *CCW/GGE.1/2018/3*, 23 October 2018. https://undocs.org/en/CCW/GGE.1/2018/3.

Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons (CCW), GGE LAWS (CCW GGE LAWS). "Report of the 2019 session of the group of governmental experts on emerging technologies in the area of lethal autonomous weapons systems." *CCW/GGE.1/2019/3*, 25 September 2019. https://documents.unoda.org/wp-content/uploads/2020/09/CCW_GGE.1_2019_3_E.pdf.

Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons (CCW), GGE LAWS (CCW GGE LAWS). "Chairperson's summary." Working paper, *CCW/GGE.1/2020/WP.7*, 19 April 2021. https://documents.unoda.org/wp-content/uploads/2020/07/CCW_GGE1_2020_WP_7-ADVANCE.pdf.

De Falco, Marco. *Stuxnet Facts Report: A Technical and Strategic Analysis*. Tallinn: NATO CCDCOE, 2012.

Defense Science Board. "Task force report: Resilient military systems and the advanced cyber threat." *United States Department of Defense*, January 2013. https://nsarchive.gwu.edu/sites/default/files/documents/2700168/Document-81.pdf.

Falliere, Nicolas, Liam O Murchu, and Eric Chien. "W32.Stuxnet Dossier version 1.4." *Symentec*, 11 February 2011. https://archive.org/details/w32_stuxnet_dossier.

Guarino, Alessandro. "Autonomous intelligent agents in cyber offence." In *5th International Conference on Cyber Conflict (CYCON 2013)*, Tallinn, Estonia: IEEE, 2013.

Heinl, Caitríona. "Artificial (intelligent) agents and active cyber defence: Policy implications." In *2014 6th International Conference On Cyber Conflict (CyCon 2014)*, edited by Pascal Brangetto, Markus Maybaum, and Jan Stinissen, Tallinn: NATO CCDCOE Publications, 2014.

Heinl, Caitríona. "Maturing autonomous cyber weapons systems: Implications for international security cyber and autonomous weapons systems regimes." In *Oxford Handbook of Cyber Security*, edited by Paul Cornish, Oxford: Oxford University Press, 2018.

Kaminska, Monica, Dennis Broeders, and Fabio Cristiano. "Limiting viral spread: Automated cyber operations and the principles of distinction and discrimination in the grey zone." In *Going Viral*, edited by T. Jančárková, L. Lindström, G. Visky, and P. Zotz, 59–72, Tallinn: NATO CCDCOE Publications, 2021.

Kovach, Christopher M. "Beyond Skynet: Reconciling increased autonomy in computer-based weapons systems with the laws of war." *Air Force Law Review* 71 (2014): 231–277.

Kushner, David. "The real story of Stuxnet." *IEEE Spectrum*, 26 February 2013. https://spectrum.ieee.org/telecom/security/the-real-story-of-stuxnet.

Liivoja, Rain, Maarja Naagel, and Ann Väljataga. *Autonomous Cyber Capabilities under International Law.* Tallinn: North Atlantic Treaty Organisation (NATO) Cooperative Cyber Defence Centre of Excellence (CCDCOE), 2019.

Mačák, Kubo. "Military Objectives 2.0: The case for interpreting computer data as objects under international humanitarian law." *Israel Law Review* 48, no. 1 (March 2015): 55–80.

Margulies, Peter. "A moment in time: autonomous cyber capabilities, proportionality and precautions." In *Autonomous Cyber Capabilities under International Law*, edited by Rain Liivoja, and Ann Väljataga, 152–180, Tallinn, Estonia: NATO CCDCOE Publications, 2021.

McFarland, Tim. "The concept of autonomy." In *Autonomous Cyber Capabilities under International Law*, edited by Rain Livoja and Ann Väljataga, 12–35, Tallinn, Estonia: NATO CCDCOE Publications, 2021.

Messinger, Eric. "Is it possible to ban autonomous weapons in cyberwar?" *Just Security* (blog), 15 January 2015. https://www.justsecurity.org/19119/ban-autonomous-weapons-cyberwar/.

Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War.* New York: W.W. Norton & Company, 2018.

Schmitt, Michael N., and Liis Vihul. "International cyber law politicized: The UN GGE's failure to advance cyber norms." *Just Security* (blog), 30 June 2017. https://www.justsecurity.org/42768/international-cyber-law-politicized-gges-failure-advance-cyber-norms/.

Schmitt, Michael N., and Liis Vihul (eds). *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations.* 2nd ed., Cambridge: Cambridge University Press, 2017.

Singer, Tassilo V.P. "Update to revolving door 2.0: The extension of the period for direct participation in hostilities due to autonomous cyber weapons." In *2017 9th International Conference on Cyber Conflict (CyCon)*, 1–13, Tallinn: IEEE, 2017. http://ieeexplore.ieee.org/document/8240332/.

Song, Jia and Jim Alves-Foss. "The DARPA cyber grand challenge: A competitor's perspective." *IEEE Security & Privacy* 13, no. 6 (2015): 72–76.

Tammet, Tanel. "Autonomous cyber defence capabilities." In *Autonomous Cyber Capabilities under International Law*, edited by Rain Livoja and Ann Väljataga, Tallinn, Estonia: NATO CCDCOE Publications, 2021.

Tattersall, Alec, and Damian Copeland. "Reviewing autonomous cyber capabilities." In *Autonomous Cyber Capabilities under International Law*, edited by Rain Liivoja and Ann Valjataga, Tallinn, Estonia: NATO CCDCOE Publications, 2021.

United States Department of Defense. "Autonomy in weapons systems", *Department of Defense Directive 3000.09*, 21 November 2012.

UNIDIR. "Side event on cyber and autonomous weapons." 14 October 2015. https://www.un.org/disarmament/update/side-event-on-cyber-and-autonomous-weapons/.

UNIDIR. "The weaponization of increasingly autonomous technologies: Autonomous weapon systems and cyber operations." 2017. https://unidir.org/publication/weaponization-increasingly-autonomous-technologies-autonomous-weapon-systems-and-cyber.

UN Office for Disarmament Affairs (UNODA). "Developments in the field of information and telecommunications in the context of international security." *United Nations*, July 2019. https://unoda-web.s3.amazonaws.com/wp-content/uploads/2019/07/Information-Security-Fact-Sheet-July-2019.pdf.

United Nations General Assembly (UNGA). "Developments in the field of information and telecommunications in the context of international security." *United Nations, A/RES/53/70*, 4 January 1999. http://undocs.org/A/RES/53/70.

United Nations General Assembly (UNGA). "Report of the group of governmental experts on developments in the field of information and telecommunications in the context of international security." *United Nations, A/65/201*, 30 July 2010. https://undocs.org/A/65/201.

United Nations General Assembly (UNGA). "Report of the group of governmental experts on developments in the field of information and telecommunications in the context of international security." *United Nations, A/68/98*, 24 June 2013. https://undocs.org/A/68/98.

United Nations General Assembly (UNGA). "Report of the group of governmental experts on developments in the field of information and telecommunications in the context of international security." *United Nations, A/70/174*, 22 July 2015. https://undocs.org/A/70/174.

United Nations General Assembly (UNGA). "Chair's summary of the open-ended working group on developments in the field of information and telecommunications in the context of international security." *United Nations, A/AC.290/2021/CRP.3*, 10 March 2021. https://front.un-arm.org/wp-content/uploads/2021/03/Chairs-Summary-A-AC.290-2021-CRP.3-technical-reissue.pdf

United Nations General Assembly (UNGA). "Final substantive report of the open-ended working group [OEWG] on developments in the field of information and telecommunications in the context of international security." *United Nations,*

A/AC.290/2021/CRP.2, 10 March 2021. https://front.un-arm.org/wp-content/uploads/2021/03/Final-report-A-AC.290-2021-CRP.2.pdf.

United Nations General Assembly (UNGA). "Report of the group of governmental experts on developments in the field of information and telecommunications in the context of international security." *United Nations, A/76/135*, 14 July 2021. https://front.un-arm.org/wp-content/uploads/2021/08/A_76_135-2104030F-1.pdf.

Wallace, David. *Cyber Weapon Reviews under International Humanitarian Law*. Tallinn: NATO CCDCOE, 2018.

9 Advanced artificial intelligence techniques and the principle of non-intervention in the context of electoral interference

A challenge to the "demanding" element of coercion?

Jack Kenny

The nature of foreign interference in the affairs of another state in the current era is fraught with complexities. Globalisation and advances in technology have enabled more opportunities and methods by which to interfere in the domestic affairs of foreign states. While traditional news media was controlled by domestic news organisations through broadcast and print, the availability of personal computers and smart mobile devices have revolutionised access to information and how news is consumed. The emergence and popularity of social media platforms on which users are able to publish their own content and share links to news articles has enabled state-sponsored disinformation campaigns to exert an unprecedented impact on public discourse in other states, creating "echo-chambers" where users are exposed to disinformation. The algorithms of online platforms repeat and reinforce those views, exacerbating extreme political positions.[1]

Beyond traditional influence operations, deepfakes of prominent public figures have prompted much discussion about the role of advanced artificial intelligence techniques in disinformation campaigns.[2] Not only have advances in technology and the invention of the internet amplified efforts to influence an electorate without the consent of the target state, technology has also lowered the technical barrier of capabilities required for conducting such operations and effectively rebalanced the traditional power dynamic of certain states in their ability to influence others. For example, democratic states with less restrictions on access to online platforms who support the freedom of expression are more susceptible to electoral interference on online platforms than states who have tighter controls or restrictions over online access and the sharing of information, and who do not hold free and fair elections. The development and availability of open-source artificial intelligence software for content creation has removed barriers relating to expertise and hardware requirements, enabling anyone with time and access to cloud computing to

DOI: 10.4324/9781003284093-12

run sophisticated machine learning processes and graphical rendering.[3] The technology for advanced artificial intelligence techniques is improving at a rapid pace, for example, experts predict that deepfake videos may soon be indistinguishable from genuine videos.[4] As technology advances and the means to use advanced artificial intelligence techniques become more widely available, their use in online disinformation campaigns is expected to become increasingly prevalent.[5] Given the purported effectiveness of disinformation campaigns on online platforms and the difficulties faced by governments and social media companies in how best to tackle such content, advanced artificial intelligence techniques hold the potential to be incredibly effective and influential in spreading disinformation, further enhancing the ability of foreign actors to interfere in the domestic affairs of states.

The principle of non-intervention is an established rule of customary international law that is the corollary of the principle of sovereignty.[6] Closely linked with the concept of a state's domestic affairs, also known as *domaine réservé*, and the international legal limits on a state's jurisdiction,[7] it prohibits a state from intervening by coercive means in matters within another state's sovereign powers.

There is currently widespread agreement that, in principle, existing international law applies to state cyber operations.[8] Reports of the UN Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (GGE), adopted by consensus, specifically confirm that the principle of non-intervention applies to state cyber operations.[9] Statements by a number of states on the application of international law to cyber operations further support the application of the principle of non-intervention to cyber operations,[10] though statements that have been published to date remain general and broad in nature.

This chapter seeks to examine the application of the principle of non-intervention in international law in relation to the use of advanced artificial intelligence techniques in disinformation campaigns of electoral interference. As artificial intelligence technology advances, techniques such as deepfakes possess characteristics that have the potential to be extremely influential in interfering in the affairs of another state by means of deception in various forms. This chapter uses these techniques as a case study to challenge and explore the limits of the coercion element required to establish a prohibited intervention. In the context of cyber operations, it has been argued that the non-intervention principle in international law involves a relatively high threshold of violation because of the requirement for the behaviour on the part of the perpetrating state to be coercive in nature.[11] This chapter makes the argument that the non-intervention principle is capable of a broader application where coercion requires a "degree of pressure" to deprive the target state of control of its state functions. While the use of sophisticated technological methods such as advanced artificial intelligence techniques in influence operations do not change or alter the criteria on which coercion is evaluated, they amplify the scale of effects that are possible to achieve from

indirect interference operations which increases the likelihood that such activity could interfere with a democratic state's inherent right to run free and fair elections and may therefore constitute a prohibited intervention. In that sense, the chapter challenges the understanding that mere influence alone is not capable of being coercive and argues that influence operations are capable of constituting a prohibited intervention.

This chapter consists of three substantive sections. The first section will introduce the principle of non-intervention and discuss its status as a rule of customary international law, including the identification of elements required for a prohibited intervention, consideration of the scope of what is encompassed within a state's *domaine réservé*, a discussion of common terminology and a brief overview of state practice that has frequently been at odds with the rule. The second section of the chapter will attempt to define the element of coercion and its application in the cyber context, including an overview of sources and debates in both non-cyber and cyber specific literature and an examination of the position of states, and the relevance of intent and outcome in such scenarios. The third section will discuss these arguments in relation to electoral interference and the challenges presented by advances in artificial intelligence techniques that are increasingly prevalent in enhancing the effectiveness of online disinformation campaigns. The chapter will then offer conclusions on its content and analysis.

The principle of non-intervention

The principle of non-intervention[12] is an established rule of international law that has been confirmed by the International Court of Justice (ICJ) to be 'part and parcel of customary international law'.[13] The principle of non-intervention is an inter-state doctrine, and as such does not apply to the activities of non-state actors unless the activities of such actors can be attributed to a state.[14] For the UN and UN member states acting through its organs, the prohibition of intervention in essentially domestic matters is set out in Article 2 (7) of the UN Charter.[15] There is a close relationship between the principle of non-intervention and the use of force, though '[w]hile the customary rules of international law relating to intervention have now to a considerable extent to be considered alongside the more general prohibition on the use of force, intervention is still a distinct concept.'[16] While the principle of non-intervention is recognised as part of customary international law,[17] the principle has been described in the literature as vague and 'elusive'.[18] In particular, while the prohibition applies to both interventions by force and non-forcible interventions, the application of the rule is not clearly defined outside of scenarios involving use of force,[19] where 'much depends on context, and even on the state of relations between the states concerned'.[20]

In the *Nicaragua* case the ICJ confirmed that the non-intervention principle is 'part and parcel of customary international law', while also recognising that 'examples of trespass against this principle are not infrequent'.[21] The Court

further elaborated on the legal effect of states' non-compliance with the principle, explaining that '[i]t is not to be expected that in the practice of States the application of the rules in question should have been perfect, in the sense that States should have refrained, with complete consistency, from the use of force or from intervention in each other's internal affairs.'[22] The Court also noted that '[e]xpressions of an *opinio juris* regarding the existence of the principle of non-intervention in customary international law are numerous and not difficult to find.'[23] Lowe notes the difficulty in reconciling the pervasive practice of intervention with the non-intervention rule.[24]

While states frequently condemn the acts of other states as intervention in their internal affairs,[25] there is clearly a 'long-standing contrast between the word and deeds', where 'non-intervention is preached, but not practiced'.[26] In this sense, state practice and *opinio juris* concerning the non-intervention principle are often at odds with each other,[27] with the frequent breach of the principle even leading some to question its status as law.[28]

Elements of a prohibited intervention

In the *Nicaragua* case, the ICJ identified two elements that are required for an unlawful intervention:

> A prohibited intervention must … be one bearing *on matters in which each State is permitted, by the principle of State sovereignty, to decide freely.* One of these is the choice of a political, economic, social and cultural system, and the formulation of foreign policy. Intervention is wrongful when it uses *methods of coercion in regard to such choices*, which must remain free ones. The element of coercion, which defines, and indeed forms the very essence of, prohibited intervention is particularly obvious in the case of an intervention which uses force, either in the form of military action, or in the indirect form of support for subversive or terrorist armed activities within another State.[29]

First, the activity must bear 'on matters in which each State is permitted, by the principle of State sovereignty to decide freely' (i.e., the *domaine réserve*, or areas where states are free from international obligations and regulation[30]), and second, it must involve 'methods of coercion in regard to such choices'.[31] In other words, to constitute a prohibited intervention, the coercive conduct in question must impinge upon 'matters in which a state is permitted to decide freely',[32] in what is known as its *domaine réservé*. The principle of non-intervention as a rule of customary international law and the required elements to establish a violation thereof derives directly from the rights and duties that comprise the principle of sovereignty, namely in relation to the protection of a state's political independence. As identified by Ziegler, '[n]on-interference in the *domaine réservé* is a fundamental right of States derived from sovereignty and protected by the principle of non-intervention in their internal affairs.'[33]

Scope of domaine reserve

A state's *domaine réservé* is a state's internal affairs, or a sphere of activity which '[is] not, in principle, regulated by international law'.[34] Largely, as a result of globalisation bringing about an increase in interdependence, the further development of international law, and an increase in international integration, the scope of a state's *domaine réservé* is considered to be increasingly limited as fewer areas remain free from international rules.[35] Although traditionally a state's *domaine réservé* may be considered to encompass the organisation of government, the treatment of citizens and their use of territory, today it is not always possible to identify entire policy areas that remain within the scope of a state's *domaine réservé* as many aspects of these matters have been significantly internationalised.[36]

Notwithstanding, in *Nicaragua* the ICJ identified the 'choice of a political system' as a clear-cut example of an area within a state's *domaine réservé*.[37] As such, cyber activities conducted by foreign states that affect either the process by which elections are conducted or their outcome may qualify as a prohibited intervention, provided the coercion element is also satisfied.[38] Given that electoral interference falls within a state's *domaine réservé*, any conclusions of this chapter related to electoral interference are dependent upon the fulfilment of the *domaine résérve* element.

"Interference" and "intervention"

The terminology of "intervention", "interference" and "coercion" are vague and often used in political rhetoric, with intervention and interference sometimes used interchangeably,[39] though "interference" suggests a wider scope of activities, especially when used alongside "intervention".[40] Oppenheim addresses the terminology of "interference" in consideration that mere interference does not amount to a prohibited intervention: 'Interference pure and simple is not intervention'.[41]

This chapter therefore uses the term "interference" to describe a wide range of activities that are not necessarily coercive and therefore may not amount to a prohibited intervention. "Intervention" is used to describe activities that involve a coercive intervention in the internal or external affairs of another state.

The element of coercion and its application in the cyber context

The coercion element delimits the principle from all lawful forms of state-to-state interaction that may have an effect on another state,[42] playing a crucial role in the purpose of the prohibition on intervention, which is 'to provide an acceptable balance between the sovereign equality and independence of states on the one hand and the reality of an interdependent world and the international law commitment to human dignity on the other'.[43] In other words,

its requirement 'regulates the line' between mere interference, which is not considered to constitute a violation of international law, and a prohibited intervention.[44] In this sense the coercion element 'defines, *and indeed forms the very essence of,* prohibited intervention'.[45]

Definition of coercion

Despite its critical role in defining a prohibited intervention, there is no precise definition for what constitutes 'coercion' in international law.[46] Parallels with the use of the term in domestic law and philosophical contexts must be regarded cautiously,[47] as the use of the term in international law has its own distinct meaning.[48] An examination of coercion as it arises in different contexts in international law reveals the term is used inconsistently in its employment by various entities.[49] Article 52 of the Vienna Convention on the Law of Treaties deals with coercion in the context of the conclusion of a treaty,[50] safeguarding against the conclusion of a treaty as a result of threat or use of force. However, the term coercion within Article 52 is limited to coercion by the threat or use of force that must be unlawful to invalidate the treaty in question.[51] Similarly, Article 18 of the International Law Commission's Articles on State Responsibility offers little assistance in defining coercion, where Article 18 involves the allocation of the responsibility that attaches to the act of the coerced state to the coercing state (if the act in question constitutes an internationally wrongful act).[52] The commentary to Article 18 makes it clear that the reference to coercion does not necessarily mean 'unlawful' coercion.[53] However, the commentary provides sparse guidance on the definition of coercion beyond stating that most instances within the scope of Article 18 will be unlawful 'because they involve a threat or use of force contrary to the Charter of the United Nations, or because they involve intervention, ie coercive interference, in the affairs of another State'.[54]

At the same time, it is generally understood that only acts of a certain magnitude are likely to qualify as coercive.[55] In the *Nicaragua* case, the Court drew significant support from the 1970 Declaration on the Principles of International Law concerning Friendly Relations and Cooperation among States in which the principle of non-intervention features prominently.[56] The text of the Declaration refers to coercion as follows:

> No State or group of States has the right to intervene, directly or indirectly, for any reason whatever, in the internal or external affairs of another State. Consequently, armed intervention and all other forms of interference or attempted threats against the personality of the State or against its political, economic and cultural elements, are in violation of international law. No State may use or encourage the use of economic, political or any other type of measures to coerce another State in order to obtain from it the subordination of the exercise of its sovereign rights and to secure from it advantages of any kind.[57]

According to Oppenheim, 'the interference must be forcible or dictatorial, or otherwise coercive, in effect depriving the state intervened against of control over the matter in question'.[58] For Jamnejad and Wood, '[o]nly acts … that are intended to force a policy change in the target state will contravene the principle',[59] though '[i]f the target state wishes to impress the intervening state and complies freely, or the pressure is such that it could reasonably be resisted, the sovereign will of the target state has not been subordinated.'[60]

The challenges posed by cyber operations have brought discussion about the content of the element of coercion to the forefront of debate among states and commentators.[61] In the literature on cyber operations, the non-intervention principle is often discussed as requiring a relatively high threshold of violation because of the requirement for the behaviour on the part of the perpetrating state to be coercive in nature.[62] Drawing on the text of the Declaration on Friendly Relations,[63] the *Tallinn Manual 2.0* group of experts consider coercion 'refers to an affirmative act designed to deprive another State of its freedom of choice, that is, to force that State to act in an involuntary manner or involuntarily refrain from acting in a particular way',[64] where 'coercion must be distinguished from persuasion, criticism, public diplomacy, propaganda, retribution, mere maliciousness, and the like in the sense that, unlike coercion, such activities merely involve either influencing (as distinct from factually compelling) the voluntary actions of the target State or seek no action on the part of the target State at all'.[65]

Schmitt, the Director of the *Tallinn Manual* projects, paraphrases the text of the *Tallinn Manual 2.0* with the narrower definition that:

> … a coercive act is one designed to compel another state to take action it would otherwise not take, or to refrain from taking action it would otherwise engage in.[66] [Coercion] is accordingly *more than mere influence*. It involves undertaking measures that deprive the target State of choice… diplomacy and *propaganda*, albeit intended to cause another State to act in a certain manner, *do not qualify as intervention because the target State retains the ability to choose; the decisions they are meant to affect remain voluntary*, even though they may now be suboptimal.[67]

Schmitt later elaborates on this position but without reference to the *Tallinn Manual 2.0*:

> Restated, an act of coercion is one that deprives another State of choice by either causing that State to behave in a way it otherwise would not or to refrain from acting in a manner in which it otherwise would act. *Merely influencing the other State's choice does not suffice*; *the choice to act or not has to effectively be taken off the table in the sense that a reasonable State in the same or similar circumstances would no longer consider it to be a viable option*.[68]

As this threshold of coercion is 'seldom reached', the editors of the *Tallinn Manual 2.0* argue that the vast majority of hostile cyber operations attributable to states would only implicate a possible violation of sovereignty.[69] As such, they propose that a rule of sovereignty be developed that acts as a "normative firewall" to prohibit certain cyber operations below this threshold[70] that is capable of covering instances of interference with 'inherently governmental functions' even in scenarios where there is no coercion.[71] Schmitt and Biller argue that the demanding element of coercion required for a prohibited intervention means that states who fail to endorse a "rule" of sovereignty as Schmitt proposes deprives states of the 'most likely legal basis' for taking counter-measures and 'affords other [s]tates the flexibility to act in an 'indiscriminate and reckless' manner while claiming to operate within the boundaries of international law'.[72] However, it is unclear why if the "rule" of sovereignty Schmitt proposes exists whereby an interference with or usurpation of 'inherently governmental functions' constitutes a violation of that "rule",[73] the ICJ and states made such an effort to identify and develop the requirement that an act be coercive to constitute a prohibited intervention. While the *Tallinn Manual 2.0* group of experts were unable to definitively define the term 'inherently governmental functions',[74] the text states that '[u]surpation of an inherently governmental function differs from intervention in that the former deals with inherently governmental functions, whereas the latter involves the *domaine réservé*, concepts that overlap to a degree but that are not identical.'[75] No further elaboration on this distinction is provided.

However, it is possible to identify support for a broader understanding of coercion in language suggesting various means and techniques whereby a state may coerce another state in relation to the exercise of the latter's state power that do not necessarily require the intervention to be limited to "forcing" a policy or government change,[76] or compelling 'another state to take action it would otherwise not take, or to refrain from taking action it would otherwise engage in' per se.[77] The Friendly Relations Declaration refers to '[securing] … advantages of any kind', and this formula is reiterated in a number of UN General Assembly resolutions and in the Charter of the Organization of American States,[78] though the phrase has been criticised for being an overstatement that is misleading in what may constitute coercive behaviour.[79] *Oppenheim*'s phrasing of 'the interference must be forcible or dictatorial, *or otherwise coercive*' implies that coercion may take other forms.[80] In the *Nicaragua* case the ICJ determined that coercive behaviour can be direct as well as indirect,[81] where intervention is wrongful 'when it uses methods of coercion in regard to such choices ... *which must remain free ones*' and provided a series of different examples that may or may not qualify as an unlawful intervention.[82] While acts involving forcible pressure may be a more clear case of coercion,[83] publicists who have sought to clarify the content of coercion often envision a spectrum of action.[84] As noted by Damrosch, '[t]he traditional formulation of intervention as "dictatorial interference" resulting in the "subordination of the will" of one sovereign to another is... unsatisfactory, because some subtle

techniques of political influence may be as effective as cruder forms of domination'.[85] Kunig considers that intervention 'aims to impose certain conduct of consequence on a sovereign state'.[86] Higgins warns against the oversimplification of linear notions based on the invasiveness of acts in determining the coercive nature of an act, because not all maximally invasive acts are unlawful and not all minimally invasive acts are lawful.[87] Joyner considers coercion merely 'involves … compelling the government of another State to think or act in a certain way by applying various kinds of pressure, threats, intimidation or the use of force.'[88]

Upon closer examination, coercion is capable of a broader definition that may be understood as behaviour aimed at seeking an advantage of some kind by depriving the target state of its free will over the exercise of its sovereign powers. Moynihan argues that coercion may only require a degree of pressure to deprive the target state of control of its state functions,[89] where 'coercive behaviour may extend beyond forcing a change of policy to other aims, such as preventing the target state from implementing a policy or restraining its ability to exercise its state powers in some way'.[90] While the *Tallinn Manual 2.0* commentary examples of non-intervention involve clear instances of a particular policy decision being forced on the target state,[91] its definition of coercion appears to support an understanding that 'could include restraining a state from exercising its state functions more broadly, as well as forcing it to act in a particular way':[92] '[Coercion] refers to an affirmative act *designed to deprive another State of its freedom of choice*, that is, to force that State *to act in an involuntary manner or involuntarily refrain from acting in a particular way*'.[93] For some of the *Tallinn Manual 2.0* group of experts, 'to be coercive it is enough that an act has the effect of depriving the State of control over the matter in question'.[94] This suggests coercion may be satisfied in a wider range of scenarios such as operations that aim to merely disrupt or undermine the ability of another state to exercise control over its sovereign functions.[95] In Schmitt's restatement of the definition of coercion in the *Tallinn Manual 2.0*, the quote contains a broader reference to 'measures that deprive the target state of choice'.[96] '[M]easures that deprive the target state of choice' is far broader than an act 'designed to compel another state to take action it would otherwise not take, or to refrain from taking action it would otherwise engage in', or where 'the choice to act or not has to effectively be taken off the table in the sense that a reasonable State in the same or similar circumstances would no longer consider it to be a viable option'.[97] For Watts, 'actions merely restricting a state's choice with respect to a course of action or compelling a course of action may be sufficient to amount to violations of the principle of non-intervention'.[98] Buchan simply states that coercion 'subordinates the will of the state in order for the entity exercising coercion to realise certain objectives'.[99]

Alternatively, several authors have argued that the principle of non-intervention, usually the coercion element itself, should be circumvented or weakened to have a cyber-specific application.[100] Despite these proposals, it

is unclear why the same criteria to assess coercion should not apply equally in both the cyber and non-cyber context,[101] unless state practice and *opinio juris* results in the development of customary international law that determines otherwise. As authors making such assertions provide no state practice or *opinio juris* in support of their arguments they constitute *lex ferenda*. Others have suggested that in certain cyber operations that involve electoral interference the principle of self-determination may be more relevant in establishing an internationally wrongful act, avoiding the need to satisfy the demanding coercion element required for a prohibited intervention.[102]

Positions of states in the cyber context

Among the limited number of states that have made statements on how they consider international law to apply to cyber operations, several directly address the definition of coercion. Positions range from recognising a demanding element of coercion to more broad interpretations. Norway adopts a position close to that of Schmitt's narrower restatement of the text of the *Tallinn Manual 2.0*, whereby coercion 'compel[s] the target State to take a course of action, whether by act or omission, in a way that it would not otherwise voluntarily have pursued', where 'cyber activities that are merely influential or persuasive will not qualify as illegal intervention'.[103] Similarly, for the Netherlands, '[i]n essence [coercion] means compelling a state to take a course of action (whether an act or an omission) that it would not otherwise voluntarily pursue.'[104] Romania takes a similar position, whereby coercion '[means] meaning that the goal of the intervention must be to effectively change the behaviour of the target State'.[105] Australia recognises a broader understanding of coercion, where '[c]oercive means are those that effectively deprive the State of the ability to control, decide upon or govern matters of an inherently sovereign nature'.[106] Germany adopts a position that is broader still, where '[c]oercion implies that a State's internal processes regarding aspects pertaining to its *domaine réservé* are significantly influenced or thwarted and that its will is manifestly bent by the foreign State's conduct'.[107] New Zealand also adopts a broad approach, whereby '[c]oercion can be direct or indirect and may range from dictatorial threats to more subtle means of control'.[108] For Switzerland, '[t]he distinction between exerting influence, which is permissible, and coercion, which is not, must be determined on a case-by-case basis'.[109] A number of states including the UK and Australia also appear to embrace a broader understanding of coercion given the examples they provide of cyber scenarios that may constitute a violation of the principle of non-intervention.[110]

Several of these states recognise that the definition of coercion is yet to crystallise, for example, Germany recognises that the coercion element 'requires further clarification in the cyber context',[111] while the Netherlands notes generally that '[t]he precise definition of coercion, and thus of unauthorised intervention, has not yet fully crystallised in international law'.[112]

Intent and outcome

The covert nature of cyber operations raises questions about the extent to which a state's intent or motivations, and the outcome of the activity in question, may be relevant in assessing whether an act is coercive.

Literature discussing coercion in the context of cyber operations recognises that states will usually have a goal in carrying out the action in relation to the exercise of its sovereign functions.[113] However, while coercive acts will by nature involve an intention to compel an outcome or conduct, the intention of the intervening state does not necessarily need to match that of the party receiving assistance,[114] and the motive of the intervening state is not considered to be an important factor in determining a violation of the non-intervention principle.[115] Nonetheless, the coercive behaviour on the part of the perpetrating state will be intentional by nature, and intent is therefore a further constitutive element required for a violation of the non-intervention principle.[116]

It is generally understood that whether or not a coercive cyber operation produces the desired outcome has no bearing on whether there has been a violation of the non-intervention principle,[117] though the effects of the act are relevant because of the 'close causal link between the coercive behaviour and its actual or potential effects on the target state's free will to exercise control over its sovereign functions.'[118]

Some authors have argued that due to the unique nature of cyber operations greater weight should be placed on the actual or potential effects of the coercive activity on the target state's inherently sovereignty functions.[119] The majority of the *Tallinn Manual 2.0* group of experts considered that it is not required that the target state has knowledge of the coercive act.[120] For example, if a cyber operation caused an effect that disabled a target state's infrastructure in a way that caused a "degree of pressure" depriving the target state of control of its state functions, but that target state was unaware a cyber operation had caused the malfunction, perhaps believing it to be a system failure, knowledge of the coercive act is not necessary for there to be a prohibited intervention.

Electoral interference and challenges presented by artificial intelligence

In the cyber context, the non-intervention principle is frequently discussed in relation to political interference of a state in the affairs of another in relation to internal electoral processes. A number of states have directly referred to electoral interference in relation to the principle of non-intervention in the context of cyber operations.[121] As with political interference generally in the non-cyber context,[122] the conduct in question may encompass acts of greatly differing intensity and coerciveness. Perhaps the most coercive form of political interference is regime change, which has been recognised as a

clear violation of the principle of non-intervention.[123] Outside interventions involving the use of force, regime change may be achieved through the provision of support and funding for insurrectionary opposition groups, as was examined in the *Nicaragua* case which primarily concerned the US funding and supporting a political opposition to a foreign government.[124]

The support of a non-insurrectional political party in a foreign electoral process is understood to be less clear in application, although the element of coercion remains the 'key test' in determining whether there has been a breach of the non-intervention principle.[125] In the cyber context, online disinformation operations often provide support for candidates which are legitimately within the official electoral process. For Jamnejad and Wood, in the context of political interference, the provision of support to a party on the eve of an election is a more intrusive act because it is more likely to result in a change in government.[126] State practice of electoral interference is extensive, perhaps because political support leading up to an election may be most effective in influencing its outcome.[127] At the same time, numerous General Assembly Resolutions condemning interference in electoral processes suggests that states are particularly concerned about this form of intervention.[128]

Distinction between direct and indirect interference

For the following discussion it is helpful to draw a distinction between direct interference with election infrastructure, such as a cyber operation targeting voting systems to disrupt or alter the results of an election, and indirect interference in influencing or manipulating voter behaviour, which may consist of disinformation campaigns on online platforms.[129] In *Nicaragua* the Court found that the principle of non-intervention 'forbids all States or groups of States to intervene *directly or indirectly* in internal or external affairs of other States'.[130] The Court determined that the US funding of the contras constituted a prohibited intervention despite the fact that a number of steps took place after the US transfer of funds and before the coercive act of the Sandinista regime of Nicaragua occurred, demonstrating that coercive behaviour can be direct as well as indirect. In that sense, the Court 'adopted a more nuanced approach to understanding both coercive behaviour and intervention than simply direct, dictatorial behaviour.'[131]

Interference that involves the direct targeting of election infrastructure may, for instance, involve a cyber operation that alters election results, or that alters the status of voters or removes their eligibility to vote, or that renders all votes uncountable or systems unavailable to perform the voting process.[132] Examples might include a Distributed Denial of Service attack that renders the systems required to count or process votes temporarily unavailable, resulting in a situation where it is not possible for votes to be cast. As the perpetrating state is attempting to alter the results of the election in a direct manner to compel the target state to achieve an outcome, this is likely to amount to coercion and constitute a prohibited intervention.[133] A number of states,

including Australia,[134] Germany,[135] Israel,[136] New Zealand,[137] Norway,[138] Romania,[139] the UK,[140] and the US,[141] have identified the direct targeting of infrastructure to disrupt or alter data on those systems as a likely violation of the non-intervention principle. As Brian Egan noted while serving as US Department of State Legal Adviser, 'a cyber operation by a State that interferes with another country's ability to hold an election or that manipulates another country's election results would be a clear violation of the rule of non-intervention.'[142]

While direct interference with election infrastructure is clearly capable of being coercive in nature, and so may constitute a prohibited intervention, a more complicated issue is whether indirect interference that seeks to influence or manipulate voter behaviour may be coercive, and thus constitute a prohibited intervention. In discussing the limits of the element of coercion and the principle of non-intervention, some scholars have briefly referred to advanced artificial intelligence techniques such as deepfakes as an example of a particularly problematic method of interference.[143] The development and use of technology such as advanced artificial intelligence techniques do not alter the criteria by which coercion is examined. However, the challenge posed by such technology is whether its use in disinformation and micro-targeting, if carried out extensively by the perpetrating state, could through influence alone amount to coercion in undermining the target state's ability to conduct a free and fair election to decide its government.[144] Drawing on technical reports on the use of advanced artificial techniques in relation to electoral interference it is possible to identify several ways in which such techniques may play a significant role in electoral interference, enhancing the effectiveness of spreading disinformation on online platforms.[145]

Influencing an electorate and the use of artificial intelligence techniques

Indirect interference operations that seek to influence voter behaviour may involve spreading disinformation on social media platforms where false or misleading articles are circulated by actors using fake bot accounts. Artificial intelligence technology allows actors to create numerous accounts, using machine learning to build fake user profiles and profile photos from user data, which can then be controlled as bots en masse to give the impression that many citizens of a state share a particular view or opinion.[146] This can be used to create a "herd mentality" on online platforms during an election cycle, enabling foreign actors to have a far more pervasive effect in influencing voter behaviour than if these profiles and their activity had to be created and maintained manually.[147] The use of advanced artificial intelligence techniques such as deepfakes, that is, the creation of synthetic videos that closely resemble real videos,[148] allow foreign actors to release and spread videos containing false statements or actions of politicians, which may be used to incite political action or discredit an individual or political party. One of the reasons deepfakes are so

effective in influencing public opinion is their ability to harness confirmation bias.[149] There are already several examples of this in practice. For example, in 2019 in Malaysia, a video purporting to show the Minister for Economic Affairs engaging in a sexual encounter caused significant reputational damage.[150] In 2020, the Oxford Internet Institute identified 81 states that use online platforms to spread computational propaganda and disinformation,[151] demonstrating the extensive nature of state practice in this area.

According to the position of some who maintain a narrow or demanding understanding of the coercion element, influencing voter behaviour through disinformation campaigns may not constitute a coercive act because although such activity may be 'a noxious form of influence', 'voters (the state) retain their ability to decide for whom to vote'.[152] According to such views, merely influencing a state's choice is not coercive as it does not force that state to act in an involuntary manner or involuntarily refrain from acting in a particular way.[153] In relation to these views, for Switzerland, '[t]he distinction between exerting influence, which is permissible, and coercion, which is not, must be determined on a case-by-case basis',[154] and for Norway 'cyber activities that are merely influential or persuasive will not qualify as illegal intervention.'[155]

However, the fact that an operation seeks to compel an outcome through influence alone does not necessarily preclude it from compelling a state to act in an involuntary manner or to involuntarily refrain from acting in a particular way if that state is deprived of control over its ability to hold a free and fair election. The use of advanced artificial intelligence techniques to create an impression of a narrative among an electorate on online platforms that discredits a candidate or political party pushes the boundaries of the interpretations of coercion as outlined in this chapter. Globalisation and advances in technology constitute significant changes since the period of major UN activity on the principle of non-intervention and these developments and challenges will leave their mark on the application of the principle.[156] The effects that are possible to achieve through the use of technology such as advanced artificial intelligence techniques are unprecedented in their speed, effectiveness and scale in manipulating the behaviour of an electorate to an extent that simply wasn't possible before.[157] This chapter has demonstrated that coercion is capable of a broader application that relies on a degree of pressure to deprive the target state of control of its state functions, under which certain influence operations may amount to coercion. Accordingly, as stated by Tsagourias, 'the use of deep fakes when, during an electoral campaign, imageries, voices, or videos of politicians are simulated in order to discredit them … [t]o the extent that such operations are designed and executed in such a way as to manipulate the cognitive process where authority and will are formed and to take control over peoples' choices of government … would constitute intervention.'[158]

The question is therefore not whether influence operations may be coercive, but instead what factors or conditions are relevant in determining in which scenarios they may amount to coercion. The provision of information

on the eve of an election is a more intrusive act and the falseness of information may indicate a higher likelihood that this activity may interfere with a democratic state's inherent right to hold free and fair elections.[159] In consideration of whether propaganda may be coercive, Murty asks whether the audience's choice of alternatives have been severely restricted as a result of the operation,[160] which implies a focus on outcome, actual or expected, rather than means or methods employed.[161] For Watts '[o]nly where propaganda could be said to constitute significant support for coercion on a level commensurate with the logistical and financial support recognized by the *Nicaragua* Court, could cyber propaganda amount to a violation of the principle of non-intervention.'[162] For Jamnejad and Wood, whether the dissemination of propaganda violates the non-intervention principle depends on all the circumstances at hand: If the information that is shared is factual and neutral, then it is not likely to constitute a breach of the non-intervention principle.[163] Conversely, if the information is disinformation, that is, false or not factually accurate, such as the discussed uses of advanced artificial intelligence techniques, the likelihood that this activity may interfere with a democratic state's inherent right to hold free and fair elections is higher.[164]

Influence operations that result in disruption of election process

It is also possible that the spreading of disinformation that may include the use of advanced artificial techniques in influence operations could, though not directly targeting the infrastructure of an election, nonetheless disrupt the election process in a way that would significantly alter the outcome or results of an election, tantamount to a direct operation that targets and disrupts or alters election infrastructure systems.[165] Imagine a scenario where, on the eve of an election, an incumbent head of state appeared to announce that the election was cancelled or postponed, or that the individual was conceding or withdrawing from the race, the deepfake video of which spreads virally online resulting in a significant number of voters not casting their vote on the day of the election. While not a direct attack on voting systems to alter or disrupt those systems, such a scenario may still produce the same effect. In this case, as the perpetrating state is attempting to alter the results of the election in a manner that deprives the target state of control over its ability to conduct a free and fair election to decide its government, equivalent to the direct electoral interference highlighted by several states as an example of a prohibited intervention.[166] Such a scenario would likely to amount to coercion and constitute a prohibited intervention.[167]

Individual influence operations as part of a broader influence campaign

While foreign interference in the 2016 US elections has received significant attention from scholars of international law on questions relating to the

principle of non-intervention,[168] it nonetheless remains a particularly useful case study to demonstrate several issues relating to indirect electoral interference that concern influencing or manipulating voter behaviour within the context of a larger influence campaign. The influence operation had two main components: The exfiltration and publication of the Democratic National Committee materials, and the spreading of disinformation on social media to influence voter choice.[169]

Regarding the first component of the exfiltration and publishing of material, the material was verified as authentic,[170] and many subsequent news articles based on its publication were both factual and in the public interest. That this step may constitute a prohibited intervention is not obvious,[171] because the published material is factual it is less likely to be coercive and constitute a breach of the non-intervention principle.[172] However, if a scenario took place whereby sensitive state secrets or classified documents were exfiltrated by state actors who threatened to publish or disclose said documents unless the state changed their policy on a particular matter within its *domaine réserve*, that action may be coercive and thus would constitute a prohibited intervention.

For the second element of spreading falsified news on online platforms to influence the election results, the information was false and not factually accurate and was disseminated on the eve of an election,[173] so the likelihood that this activity may be coercive in interfering with a democratic state's inherent right to hold free and fair elections is higher[174] and depending on these factors and the definition of coercion outlined above, the operation may constitute a prohibited intervention.[175]

Some authors simply state that it is unclear whether the operations were coercive and therefore that they did not constitute a prohibited intervention,[176] while others maintain it violated the principle based on a broad or modified interpretation of coercion.[177] In relation to those that do not consider the operations to be coercive, Schmitt recognises what he considers to be 'a slightly sounder view', that 'the cyber operations manipulated the process of elections and therefore caused them to unfold in a way that they otherwise would not have', and so may be coercive.[178] This position appears to contradict his assertion in the same paper that '… diplomacy and propaganda, albeit intended to cause another State to act in a certain manner, do not qualify as intervention *because the target State retains the ability to choose; the decisions they are meant to affect remain voluntary*, even though they may now be suboptimal'.[179] The operations constituted indirect interference as they did not affect the state or the electorate's ability to choose a candidate during the election in so far as the decisions of the electorate remained voluntary. This debate demonstrates the difficulty in maintaining the position that influence operations alone are not capable of amounting to coercion, which will only be exacerbated by the increased use and effectiveness of advanced artificial intelligence techniques in future influence operations.

While in practice it is difficult to separate individual influence operations from larger campaigns of influence, Schmitt has suggested that the scale and

effects test from the ICJ's *Paramilitary Activities* judgment assessing whether a use of force rises to threshold of an armed attack might be adopted in the context of intervention, and cites the recent position of Germany[180] which he suggests supports this proposal.[181] However, despite Schmitt's claim that '[t]he German adoption of the approach in the context of intervention is further evidence that it is gaining widespread acceptance as a means of assessing international law thresholds more generally when applied to cyber operations', which he considers to be 'an appropriate use of the scale and effect standard',[182] there is sparse evidence in literature and among states supporting the assertion that this particular standard that the ICJ formulated in the context of use of force is directly transposable to the principle of non-intervention. Rather, the statement by Germany is better understood as a general statement recognising that individual cyber operations often form part of a larger campaign of influence that may be coercive as a whole and thus may constitute a prohibited intervention. It remains to be seen whether more states will recognise such an approach in light of the further development and increased availability of technology such as advanced artificial intelligence techniques and their ability to enhance the effectiveness of disinformation campaigns.[183]

Conclusion

The element of coercion plays a key role that 'regulates the line' between mere interference, which is not generally considered to constitute a violation of international law, and a prohibited intervention.[184] The coercive behaviour on the part of the perpetrating state will be intentional by nature, and intent is a further constitutive element required to establish a prohibited intervention. Though the ICJ has recognised the principle of non-intervention is a rule which has been frequently breached, it confirmed this has not affected its status as a rule of custom, and that it remains 'part and parcel of customary international law.'[185]

Some commentators in the literature on the application of international law to cyber operations have asserted that the element of coercion sets a demanding standard that is seldom met. However, this chapter argues that the non-intervention principle is capable of a broader application where coercion requires the exertion of a "degree of pressure" to deprive the target state of control of its state functions, where coercive behaviour may extend beyond forcing a change of policy to preventing the target state from implementing a policy or restraining its ability to exercise state powers. Indeed, several states have recently recognised broader and more nuanced understandings of coercion under which influence operations may qualify as prohibited interventions.[186] However, it is important to note that to date no state has explicitly claimed a cyber operation has violated the principle of non-intervention.

The primary focus of the application of the rule of non-intervention to cyber operations among states and commentators concerns electoral interference. While several states now recognise the application of the principle

of non-intervention to scenarios of electoral interference that directly target election infrastructure to disrupt or alter the result of an election,[187] a more complicated issue is whether indirect operations that seek to influence or manipulate voter behaviour may be coercive, and thus constitute a prohibited intervention.

Advances in technology increasingly enable indirect influence operations to be capable of amounting to coercion insofar as the have the ability to undermine the target state's ability to conduct a free and fair election to decide its government, in effect applying a "degree of pressure" to deprive the target state of control of its state functions. Developments in technology such as the use of advanced artificial intelligence techniques in influence operations do not change or alter the criteria on which coercion is evaluated. However, they amplify the scale of effects that are possible to achieve from influence operations which increases the likelihood that such activity could interfere with a democratic state's inherent right to run free and fair elections. As such, they increasingly push the boundaries of indirect interference operations that influence or manipulate voter behaviour, and methods utilising such techniques are capable of being coercive and may therefore constitute a prohibited intervention.

In that sense, the continued development of technology such as advanced artificial intelligence techniques that enhance the effectiveness of disinformation operations may be said to constitute a challenge to the traditional understanding that mere influence alone is not coercive and therefore may not qualify as a prohibited intervention. Accordingly, the question is therefore not whether influence operations may be coercive, but instead what factors or criteria are relevant to consider in determining in which scenarios they may amount to coercion. Individual cyber operations or disinformation operations have the ability to be effective in influencing an electorate as a whole, and states have begun to discuss whether individual operations which may not themselves be coercive may collectively amount to coercion and thus be capable of constituting a prohibited intervention. As in practice it is difficult to distinguish between individual influence operations that are part of a broader campaign of activity, it seems logical that cumulative effects of individual operations must be capable of resulting in a "degree of pressure" to deprive the target state of control of its state functions. More clarity is likely to emerge as states continue to engage in international fora and release statements outlining their positions.

Notes

1 For an exploration of these issues, see Samuel C. Rhodes, "Filter bubbles, echo chambers, and fake news: How social media conditions individuals to be less critical of political misinformation," *Political Communication* 39, no. 1 (2021).
2 See Simon Parkin, "The rise of the deepfake and the threat to democracy," *The Guardian*, 22 June 2019; Alexander Puutio and David Alexandru Timis, "Deepfake democracy: Here's how modern elections could be decided by fake news," *World Economic Forum*, 5 October 2020.

3 Puutio and Timis, "Deepfake Democracy."
4 James Rundle, "FBI warns deepfakes might become indistinguishable from reality," *Wall Street Journal*, 17 January 2020, sec. WSJ Pro.
5 While many in the US predicted deepfakes would play a significant role in the 2020 elections, actors relied on simpler forms of disinformation. However, as online platforms implement mechanisms to combat these simpler forms of disinformation and the technology to manufacture deepfake videos becomes more widely available, deepfakes are expected to play an increased role in online disinformation, see: Rachel Metz, "The fight to stay ahead of deepfake videos before the 2020 US election," *CNN*, 12 June 2019; Tom Simonite, "What happened to the deepfake threat to the election?" *Wired*, 16 November 2020; Tim Mak and Dina Temple-Raston, "Where are the deepfakes in this presidential election?" *NPR*, 1 October 2020, sec. Investigations.
6 For Oppenheim, '[i]ts prohibition is the corollary of every state's right to sovereignty, territorial integrity and political independence,' citing Military and Paramilitary Activities Case, ICJ Rep (1986), 106–107, R.Y. Jennings and Arthur Watts, *Oppenheim's International Law* (London: Longman, 1996), 428. It has moreover been presented as a corollary of the principle of the sovereign equality of States', *Military and Paramilitary Activities in and Against Nicaragua* (*Nicaragua v United States of America*) (Merits) [1986] ICJ Rep 14, 202 [hereafter *Nicaragua*].
7 Maziar Jamnejad and Michael Wood, "The principle of non-intervention," *Leiden Journal of International Law* 22, no. 2 (June 2009): 346.
8 Three reports of the UN Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (GGE), comprising experts from a number of states working in their personal capacity, have determined that international law, 'in particular the United Nations Charter', applies to cyber operations, Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (2013), UN Doc A/68/98, 8; Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (2015), UN Doc A/70/174, 12; Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security (2019–2021), UN Doc A/76/135, 17. The 2015 report was subsequently adopted by the UN General Assembly by consensus, Resolution 70/237 (2015), UN Doc A/RES/70/237. The final report of the UN Open-Ended Working Group on developments in the field of information and telecommunications in the context of international security (OEWG), a parallel process open to all interested states, with participation from the private sector, NGOs and academia, recently reaffirmed that international law is applicable to cyber operations, Final Substantive Report of the UN Open-Ended Working Group on developments in the field of information and telecommunications in the context of international security (2021), UN Doc A/AC.290/2021/CRP.2, 2.
9 There is currently widespread agreement that, in principle, existing international law applies to state cyber operations, see 2015 GGE report, UN Doc A/70/174, 12; 2019–2021 GGE report, UN Doc A/76/135, 17.
10 See the positions of Germany, "On the application of international law in cyberspace," *The Federal Government*, March 2021: 4–6; Australia, in Australia's International Cyber Engagement Strategy, "2019 International Law Supplement to Australia's International Cyber Engagement Strategy, Annex A: Supplement to Australia's Position on the Application of International Law to State Conduct in Cyberspace," 2019; Finland, "International Law and Cyberspace, Finland's National Positions," 3–4; the Netherlands, in Tweede Kamer der Staten-Generaal,

"Letter of the Minister of Foreign Affairs to Parliament, 5 July 2019, Kamerstuk 33 694, Nr. 47, available (in Dutch)," Officiële publicatie, 5 July 2019, 3; Israel, in Roy Schöndorf, "Israel's perspective on key legal and practical issues concerning the application of international law to cyber operations," *International Law Studies* 97 (2021): 403; the UK in Attorney General Jeremy Wright, "Cyber and international law in the 21st century," *Chatham House*, 23 May 2018; and positions of the US in Brian Egan, "Remarks on international law and stability in cyberspace," *Berkeley Law School, California*, 10 November 2016 and Hon. Paul C. Ney, Jr., "DOD General Counsel Remarks at U.S. Cyber Command Legal Conference," 2 March 2020.

11 Michael N. Schmitt and Liis Vihul, "Sovereignty in cyberspace: Lex Lata Vel Non?" *AJIL Unbound* 111 (2017): 214; Ido Kilovaty, "The elephant in the room: Coercion," *AJIL Unbound* 113 (2019): 89–90; Katharina Ziolkowski, "General principles of international law as applicable in cyberspace," in *Peacetime Regime for State Activities in Cyberspace: International Law, International Relations and Diplomacy* (Tallinn, Estonia: NATO CCD COE Publication, 2013), 165.

12 The principle of non-intervention is stated in UN General Assembly Resolution 2625: 'No State or group of States has the right to intervene, directly or indirectly, for any reason whatever, in the internal or external affairs of any other State. Consequently, armed intervention and all other forms of interference or attempted threats against the personality of the State or against its political, economic and cultural elements, are in violation of international law… No State may use or encourage the use of economic, political or any other type of measures to coerce another State in order to obtain from it the subordination of the exercise of its sovereign rights and to secure from it advantages of any kind.' An early formulation of the principle was adopted by the General Assembly in 1965 as Resolution 2131(XX): 'No State has the right to intervene, directly or indirectly, for any reason whatever, in the internal or external affairs of any other State. Consequently, armed intervention and all other forms of interference or attempted threats against the personality of the State or against its political, economic and cultural elements are condemned.' The Resolution further condemns 'economic, political or any other type of measures to coerce another State', the assisting or inciting of subversive or armed activities aimed at overthrowing the state, the use of force to deprive peoples of their national identity; and emphasises the obligation to respect the principles of non-intervention and self-determination; Article 8 of the Montevideo Convention on the Rights and Duties of States 1933; Declaration on the Inadmissibility of Intervention in the Domestic Affairs of States and the Protection of Their Independence and Sovereignty, UN Doc A/RES/20/2131, 21 December 1965; Declaration on Principles of International Law Concerning Friendly Relations and Co-operation Among States in Accordance with the Charter of the United Nations, Principle C, UN Doc A/RES/25/2625, 24 October 1970; Declaration on the Inadmissibility of Intervention and Interference in the Internal Affairs of States, UN Doc A/RES/36/103, 9 December 1981.

13 *Certain Activities Carried Out by Nicaragua in the Border Area (Costa Rica v Nicaragua) (Merits) [2015] ICJ Rep 2015,* 202.

14 See Part One, Chapter II of the International Law Commission's Articles on State Responsibility.

15 See Jennings and Watts, *Oppenheim's International Law*, 430.

16 Ibid., 429.

17 As recognised by the ICJ on several occasions, see *the Corfu Channel Case* (*United Kingdom v Albania*) (Merits), ICJ Reports 1949, 4, at 34–35; *Nicaragua*, 202–205; in *Nicaragua*, where the Court referred to 'the customary-law principle of

non-intervention', *Nicaragua*, para 245; Judge Jennings stated in his Dissenting Opinion that '[t]here can be no doubt that the principle of non-intervention is an autonomous principle of customary law; indeed, it is much older than any of the multilateral Treaty regimes in question', *Nicaragua*, at 534; Citing the *Nicaragua* judgment, Oppenheim states '[t]hat intervention is, as a rule, forbidden by international law there is no doubt.' For Oppenheim, '[t]hat intervention is, as a rule, forbidden by international law there is no doubt. Its prohibition is the corollary of every state's right to sovereignty, territorial integrity and political independence', citing *Military and Paramilitary Activities* Case, 106–7, Jennings and Watts, *Oppenheim's International Law*, 428.

18 Lowe states that '[t]he most interesting question regarding the principle of non-intervention in international law is why on earth anyone should suppose that it exists', Vaughan Lowe, "The principle of non-intervention: Use of force," in *The United Nations and the Principles of International Law: Essays in Memory of Michael Akehurst* (London: Routledge, 1994), 67; in 2001 D'Amato questioned the existence of such a rule in practice, in-part based on criticism of customary international law and the Nicaragua judgment itself, see Anthony D'Amato, "There is no norm of intervention or non-intervention in international law," *International Legal Theory* 7, no. 1 (2001) and "Trashing customary international law," *American Journal of International Law* 81, no. 1 (January 1987).

19 Outside the area of use of force, 'it is often unclear what is, and what is not, prohibited under customary international law. Much depends on context, and even on the state of relations between the states concerned.' Jamnejad and Wood, "The principle of non-intervention," 367.

20 Ibid.

21 *Certain Activities*, 202.

22 *Nicaragua*, 186.

23 *Nicaragua*, 202.

24 'From the most cursory review of the international history of the past two centuries it is apparent that intervention in foreign States is quite normal. Indeed, if international history is thought of as the analysis of the influences of nations upon each other, it is arguable that the very terrain of history is mapped out on the grid of intervention...The most interesting question regarding the principle of non-intervention in international law is why on earth anyone should suppose that it exists', Lowe, "The principle of non-intervention," 66–67.

25 Jamnejad and Wood, "The principle of non-intervention," 346.

26 Lowe, "The principle of non-intervention," 72.

27 Writing in 1989, Damrosch identified 'a rather serious gap between what a broad view of the nonintervention norm would require and what states actually do,' Lori Fisler Damrosch, "Politics across borders: Nonintervention and nonforcible influence over domestic affairs," *American Journal of International Law* 83, no. 1 (January 1989): 23; also see Jamnejad and Wood, who note that '[t]he abstract rhetoric of the law, as expressed in resolutions of the UN General Assembly and other bodies, is hardly reflected in the practice of states,' in Jamnejad and Wood, "The principle of non-intervention," 349.

28 See D'Amato, "There is no norm."

29 *Nicaragua,* 202.

30 Generally, see Katja S. Ziegler, "Domaine réservé," in *Max Planck Encyclopedia of Public International Law* (Oxford University Press, April 2013).

31 Ibid.

32 *Nicaragua*, 205.

33 Ziegler, "Domaine réservé."

34 *Nationality Decrees in Tunis and Morocco*, PCIJ Rep Series B No 4 (1923) 23.

35 See Ziegler, "Domaine Réservé; Philip Kunig, "Intervention, prohibition of," in *Max Planck Encyclopedia of Public International Law*, 16 (April 2008); Jamnejad and Wood, "The principle of non-intervention," 346.
36 For example, the impact of human rights law on jurisdiction over and the treatment of foreign nationals on a state's territory, admission of foreign nationals onto a state's territory, see Ziegler, "Domaine réservé."
37 *Nicaragua*, the 'choice of political system' is a matter falling within a state's sovereign prerogatives which should remain 'free from external intervention,' 205.
38 Michael N. Schmitt, ""Virtual" disenfranchisement: Cyber election meddling in the grey zones of international law," *Chicago Journal of International Law* 9 (2018): 49.
39 See Antonios Tzanakopoulos, "The right to be free from economic coercion," *Cambridge Journal of International and Comparative Law* 4, no. 3 (2015): 623.
40 Kunig, "Intervention, prohibition of;" Kilovaty, "Elephant in the room," 167; Michael Schmitt, *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge: Cambridge University Press, 2017), 313. The terms have also been used in the context of jus ad bellum and the question of 'humanitarian intervention', see Vaughan Lowe and Antonios Tzanakopoulos, "Humanitarian intervention," in *Max Planck Encyclopedia of Public International Law* (Oxford University Press, May 2011); Jamnejad and Wood, "The principle of non-intervention," 347.
41 Jennings and Watts, *Oppenheim's International Law*, 432.
42 Jamnejad and Wood, "The principle of non-intervention," 381.
43 Rosalyn Higgins, *Themes and Theories* (Oxford: Oxford University Press, 2009), 273; Rosalyn Higgins, "Intervention and international law," in *Intervention In World Politics*, ed. Hedley Bull (Oxford: Clarendon Press, 1984), 30.
44 Jamnejad and Wood, "The principle of non-intervention," 381.
45 According to the *Nicaragua* judgment (emphasis added), para 205.
46 Helen Keller, "Friendly relations declaration (1970)," in *Max Planck Encyclopedia of Public International Law* (Oxford University Press, June 2009); Jamnejad and Wood, "The principle of non-intervention," 347;
47 For example, see Jens David Ohlin, "Did Russian cyber-interference in the 2016 election violate international law?" *Texas Law Review* 95 (2017), 1509–1591.
48 Harriet Moynihan, "The application of international law to state cyberattacks: Sovereignty and non-intervention," Research Paper. *Chatham House, International Law Programme* (December 2019): 29.
49 Tzanakopoulos, "The right to be free," 623.
50 Article 52 of the VCLT states that '[a] treaty is void if its conclusion has been procured by the threat or use of force in violation of the principles of international law embodied in the Charter of the United Nations.'
51 For further discussion, see Olivier Corten and Pierre Klein (eds), *The Vienna Conventions on the Law of Treaties—A Commentary* (Oxford University Press, 2011), 1201, 1205–1211.
52 Article 18 of the ILC's Articles on State Responsibility states that '[a] State which coerces another State to commit an act is internationally responsible for that act if: (a) the act would, but for the coercion, be an internationally wrongful act of the coerced State; and (b) the coercing State does so with knowledge of the circumstances of the act.'
53 ILC Articles on State Responsibility Commentaries, commentary to Article 18.
54 Ibid.
55 Jamnejad and Wood, "The principle of non-intervention," 348.
56 *Nicaragua*, 192 (citing General Assembly Res 2625 (XXV).
57 The *Friendly Relations Declaration,* Principle 3.

58 Jennings and Watts, *Oppenheim's International Law*, 428; see also Judge Schwebel in his Dissenting Opinion in Nicaragua: 'The essence of [such customary international law of non-intervention as there is] long has been recognized to prohibit the dictatorial intervention by one State in the affairs of another', para 98.
59 Jamnejad and Wood, "The principle of non-intervention," 348.
60 Ibid.
61 The UN GGE reports recognise the 'unique attributes' of cyber operations, 2013 GGE report, UN Doc A/68/98, 8; 2015 GGE report, UN Doc A/70/174, 7; 2019–2021 GGE report, UN Doc A/76/135, 8; as Gill observes, '[t]he principle of non-intervention is, on the one hand, a well-established rule of international law and, at the same time, one which is in some respects controversial and open to various definitions and differing interpretations, depending upon how widely or narrowly it is construed', Terry D. Gill, "Non-intervention in the cyber context," in *Peacetime Regime for State Activities in Cyberspace* (NATO CCD COE Publications, 2013), 217.
62 Schmitt and Vihul consider that 'the prohibition on intervention and the use of force … contain thresholds that are seldom reached,' Schmitt and Vihul, "Sovereignty in cyberspace," 214; Kilovaty refers to coercion as a 'narrow' standard, and as a result 'cyber operations require a more nuanced definition of nonintervention,' Kilovaty, "Elephant in the room," 89–90. According to Ziolkowski, the text of the Friendly Relations Declaration 'results in the notion that "coercion" occurs only in drastic cases of overwhelming (direct or indirect) force being put upon a State's free and sovereign decision-making process', and '[s]cholars assert that illegal coercion implies massive influence, inducing the affected state to adopt a decision with regard to its policy or practice which it would not envision as a free and sovereign state', Ziolkowski, "General principles," 165.
63 'No state may use or encourage the use of economic political or any other type of measures to coerce another state in order to obtain from it the subordination of the exercise of its sovereign rights and to secure from it advantages of any kind.' (GA Res 2625 (XXV), para 3.)
64 Schmitt, *Tallinn Manual 2.0*, 317.
65 Ibid., 318–319.
66 [citing Schmitt, *Tallinn Manual 2.0*, 318–319].
67 Michael N. Schmitt, "Grey zones in the international law of cyberspace," *Yale Journal of International Law* 42, no. 2 (2017): 8 (emphasis added).
68 Michael N. Schmitt, "Autonomous cyber capabilities and the international law of sovereignty and intervention," *International Law Studies* 96 (2020): 561 (emphasis added).
69 See Schmitt and Vihul, "Sovereignty in cyberspace," 214. In line with this position, Finland notes that '[c]ompared to a violation of sovereignty, the requirement of coercive nature and that of domaine réservé make the threshold of prohibited intervention considerably higher. This underlines the importance of continued understanding of sovereignty as not only a principle but also an independent primary rule of international law', "International Law and Cyberspace, Finland's National Positions."
70 As discussed in the previous chapter on the application of the principle of sovereignty to state cyber operations.
71 Michael Schmitt, "Cybersecurity and international law," in *The Oxford Handbook of the International Law of Global Security* (Oxford University Press, 2021), 669; Michael N. Schmitt, "The law of cyber warfare: Quo vadis?" *Stanford Law & Policy Review* 25 (2014): 276; Michael N. Schmitt, "Taming the lawless void: Tracking the evolution of international law rules for cyberspace," *Texas*

National Security Review 3, no. 3 (15 July 2020); Michael Schmitt and Jeffrey Biller, "Un-caging the bear? A case study in cyber opinio juris and unintended consequences," *EJIL Talk!* (blog), 24 October 2018; Michael N. Schmitt and Liis Vihul, "Respect for sovereignty in cyberspace," *Texas Law Review* 95 (2017): 1670; Michael N. Schmitt, "Peacetime cyber responses and wartime cyber operations under international law: An analytical vade mecum," *Harvard National Security Journal* 8 (2017): 243. For an alternative view, see Gary P. Corn and Robert Taylor, "Sovereignty in the age of cyber," *AJIL Unbound* 111 (2017); also see the position of the UK, which recognises no 'specific rule or additional prohibition for cyber conduct going beyond that of non-intervention', Foreign, Commonwealth & Development Office, "Application of International Law to States' Conduct in Cyberspace: UK Statement," 3 June 2021.

72 For example, concerning the implications of recognising the application of the principle of non-intervention, but not a rule of sovereignty to cyber operations, see Schmitt and Biller, "Un-caging the bear?; On the relationship between sovereignty and non-intervention in the cyber context generally, see Moynihan, "Application of international law to state cyberattacks," 48–51.

73 Schmitt, *Tallinn Manual 2.0*, 20.

74 Ibid., 22.

75 Ibid., 24.

76 Jennings and Watts, *Oppenheim's International Law*, 428; Jamnejad and Wood, "The principle of non-intervention," 348.

77 Schmitt, "Grey Zones," 8; also see Schmitt, "Autonomous cyber capabilities," 561.

78 Article 20, Charter of the Organization of American States (signed 1948, entered into force 1951).

79 Gill refers to the text 'the obtaining of advantages of any kind' in the Friendly Relations Declaration as 'potentially misleading and something of an overstatement' as 'all States attempt to influence other States to enter into favourable trade and economic relations, and attempt to increase their prestige and influence by means of economic, trade and cultural exchange policies and other forms of cooperation', Gill, "Non-intervention," 221–222.

80 Jennings and Watts, *Oppenheim's International Law*, 428.

81 '[T]he [non-intervention] principle forbids all States or groups of States to intervene directly or indirectly in the internal or external affairs of other States' *Nicaragua*, para 205.

82 Nicaragua, 242 (emphasis added).

83 'The element of coercion, which defines, and indeed forms the very essence of, prohibited intervention is particularly obvious in the case of an intervention which uses force, either in the form of military action, or in the indirect form of support for subversive or terrorist armed activities within another State,' *Nicaragua,* para 205.

84 Higgins, "Intervention and international law."

85 Damrosch, "Politics across borders," 5.

86 Kunig, "Intervention, prohibition of."

87 Ibid.

88 Christopher C. Joyner, "Coercion," in *Max Planck Encyclopedia of Public International Law* (Oxford University Press, December 2006).

89 'Coercive behaviour may thus be understood as pressure applied by one state to deprive the target state of its free will in relation to the exercise of its sovereign rights in an attempt to compel an outcome in, or conduct with respect to, a matter reserved to the target state,' Moynihan, "Application of International Law to State Cyberattacks," 32; see also Reisman, '[c]oercive behaviour may thus be understood as pressure applied by one state to deprive the target state of its free

will in relation to the exercise of its sovereign rights in an attempt to compel an outcome in, or conduct with respect to, a matter reserved to the target state', W.M. Reisman, *Nullity and Revision: The Review and Enforcement of International Judgments and Awards* (New Haven: Yale University Press, 1971), 839–40.

90 Moynihan, "Application of international law to state cyberattacks," 29.
91 For example, the use by one state of non-cyber coercive means to compel another state to adopt particular domestic legislation related to Internet service provider liability, or using such means to compel another state to refrain from becoming party to a multilateral treaty dealing with cyber disarmament or human rights online, Schmitt, *Tallinn Manual 2.0*, 313.
92 Moynihan, "Application of international law to state cyberattacks," 31.
93 Schmitt, *Tallinn Manual 2.0*, 317.
94 Ibid., 318.
95 Moynihan, "Application of international law to state cyberattacks," 32.
96 Schmitt, "Grey zones," 8.
97 Schmitt, "Autonomous cyber capabilities," 561.
98 Sean Watts, "Low-intensity cyber operations and the principle of non-intervention," in *Cyber War: Law and Ethics for Virtual Conflicts* (Oxford: Oxford University Press, 2019), 259.
99 Russell Buchan, *Cyber Espionage and International Law* (Oxford: Hart Publishing, 2019), 63.
100 For example, see Thibault Moulin, "Reviving the principle of non-intervention in cyberspace: The path forward," *Journal of Conflict and Security Law* 25, no. 3 (1 December 2020); Michael P. Fischerkepper, "Current international law is not an adequate regime for cyberspace," *Lawfare* (blog), 22 April 2021; Ido Kilovaty, "Doxfare: Politically motivated leaks and the future of the norm of non-intervention in the era of weaponized information," *Harvard National Security Journal* 9 (2018).
101 Moynihan, "Application of international law to state cyberattacks," 31.
102 See Nicolas Tsagourias, "Electoral cyber interference, self-determination and the principle of non-intervention in cyberspace," in *Governing Cyberspace: Behavior, Power, and Diplomacy* (Lanham: Rowman & Littlefield, 2020); and Ohlin, "Did Russian cyber-interference," 1580.
103 "Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States, UNODA, A/76/136, August 2021," 69.
104 "Appendix: International Law in Cyberspace, Letter of 5 July 2019 from the Minister of Foreign Affairs to the President of the House of Representatives on the International Legal Order in Cyberspace, Translation," 3.
105 "Official Compendium […] A/76/136," 77.
106 Australian Government, "Annex B: Australia's Position on How International Law Applies to State Conduct in Cyberspace | Australia's International Cyber and Critical Tech Engagement," 2020.
107 Though this statement is followed by following clarification: 'For example, it is conceivable that a State, by spreading disinformation via the internet, may deliberately incite violent political upheaval, riots and/or civil strife in a foreign country, thereby significantly impeding the orderly conduct of an election and the casting of ballots. Such activities may be comparable in scale and effect to the support of insurgents and may hence be akin to coercion in the above-mentioned sense. A detailed assessment of the individual case would be necessary,' "On the Application," 5.
108 New Zealand Ministry of Foreign Affairs and Trade, "The Application of International Law to State Activity in Cyberspace," 1 December 2020.
109 Ibid.

110 In 2018 the UK stated 'the practical application of the principle in this context would be the use by a hostile state of cyber operations to manipulate the electoral system to alter the results of an election in another state, intervention in the fundamental operation of Parliament, or in the stability of our financial system. Such acts must surely be a breach of the prohibition on intervention in the domestic affairs of states,' Wright, "Cyber and International Law." In 2021 the UK stated 'the use of hostile cyber operations to manipulate the electoral system in another State to alter the results of an election, to undermine the stability of another State's financial system or to target the essential medical services of another State could all, depending on the circumstances, be in violation of the international law prohibition on intervention', Foreign, Commonwealth & Development Office, "Application of International Law." In 2020, following closely in the footsteps of the UK, Australia stated 'the use by a hostile State of cyber activities to manipulate the electoral system to alter the results of an election in another State, intervention in the fundamental operation of Parliament, or in the stability of States' financial systems would constitute a violation of the principle of non-intervention', Australian Government, "Annex B."
111 Germany, "On the application," 5.
112 The Netherlands, "Appendix: International law," 3.
113 The majority of the international experts involved in the Tallinn Manual 2.0 considered that 'the coercive effort must be designed to influence outcomes in, or conduct with respect to, a matter reserved to a target state', Schmitt, *Tallinn Manual 2.0*, 317; Schmitt, ""Virtual" Disenfranchisement," 50; Schmitt, "Autonomous Cyber Capabilities," 561; also see Jennings and Watts, *Oppenheim's International Law*, 428.
114 In *Nicaragua,* the ICJ found 'in international law, if one State, with a view to the coercion of another State, supports and assists armed bands in that State whose purpose is to overthrow the government of that State, that amounts to an intervention by the one State in the internal affairs of the other, *whether or not the political objective of the State giving such support and assistance is equally far-reaching'*, 241 (emphasis added).
115 See Moynihan, "Application of international law to state cyberattacks," 32; Watts, "Low-intensity cyber operations," 268; Schmitt, ""Virtual" Disenfranchisement," 52; Intent, motive and purpose do not generally play a part in the international rules on state responsibility, see para 3 and 10 of the Commentary to Article 2 of the International Law Commission's Articles on State Responsibility.
116 See Nicaragua, para 241; In the cyber context, see Schmitt, *Tallinn Manual 2.0*, 321; Moynihan, "Application of international law to state cyberattacks," 32.
117 Schmitt, *Tallinn Manual 2.0*, 322.
118 Moynihan, "Application of international law to state cyberattacks," 33; The Tallinn Manual 2.0 gives the following example: '…a State may impose a ban on exports to another State. The latter launches highly disruptive cyber operations against the former's Ministry of Commerce in an effort to force that State to rescind the ban. Irrespective of whether rescission results, the cyber operations constitute prohibited intervention', Schmitt, *Tallinn Manual 2.0*, 322.
119 See Watts, "Low-intensity cyber operations," 256.
120 Schmitt, *Tallinn Manual 2.0*, 320.
121 See positions of Australia, "Annex B;" Germany, "On the application," 5; Israel in Schöndorf, "Israel's Perspective," 403; New Zealand, "Application of International Law;" Norway, in "Official Compendium […] A/76/136," 69; Romania, in "Official Compendium […] A/76/136," 77; Wright, "Cyber and International Law;" Foreign, Commonwealth & Development Office, "Application of

International Law;" and the US in Ney, "DOD General Counsel;" "Official Compendium […] A/76/136," 140.

122 See Quincy Wright, "Subversive intervention," *American Journal of International Law* 54, no. 3 (1960).

123 Jamnejad and Wood, "The principle of non-intervention," 368.

124 *Nicaragua,* para 205.

125 Jamnejad and Wood, "The principle of non-intervention," 368.

126 Ibid.

127 Ibid.

128 See (n) 10.

129 See Moynihan, "Application of international law to state cyberattacks," 40; Tsagourias makes a distinction between 'interference with the electoral administration, for example, interference with electoral registers to delete voters' names as well as on interference with the electoral infrastructure, for example, interference with the recording or counting of votes or the blocking of voting machines thus cancelling an election' and 'outcomes [that] can be affected not only by interfering with the electoral infrastructure but also by interfering with the process of will formation', Tsagourias, "Electoral cyber interference" (2020), 49–50. 'A helpful way to approach the issue is to distinguish election-related cyber activities that affect the State's ability to conduct an election from those that target voter attitudes', Michael N. Schmitt, "Foreign cyber interference in elections," *International Law Studies* 97 (2021): 746; also see discussion in the position of Germany, "On the Application," 4–6.

130 *Nicaragua*, 205.

131 Moynihan, "Application of international law to state cyberattacks," 30.

132 For examples of such operations, see Emily Tamkin, "Elections in Ghana marred by attempt to hack website and calls for the president to concede," *Foreign Policy* (blog), 8 December 2016; "Hackers target Ukraine's election website," *Yahoo*, 25 October 2014.

133 'Blocking voting by cyber means, such as by disabling election machinery or by conducting a distributed denial of service attack, would likewise be coercive', Schmitt, ""Virtual" Disenfranchisement," 50.

134 For Australia, 'the use by a hostile state of cyber activities to manipulate the electoral system to alter the results of an election in another State ... would constitute a violation of the principle of non-intervention', Australian Government, "Annex B."

135 Germany recognises that 'the disabling of election infrastructure and technology such as electronic ballots, etc. by malicious cyber activities may constitute a prohibited intervention, in particular if this compromises or even prevents the holding of an election, or if the results of an election are thereby substantially modified', Germany, "On the Application," 5.

136 Israel expressed agreement with the US DoD position that 'cyber operation by a State that interferes with another country's ability to hold an election or that manipulates another country's election results would be a clear violation of the rule of non-intervention,' in Schöndorf, "Israel's Perspective," 403; citing Ney, "DOD General Counsel."

137 New Zealand recognises that '[e]xamples of malicious cyber activity that might violate the non-intervention rule include: a cyber operation that deliberately manipulates the vote tally in an election or deprives a significant part of the electorate of the ability to vote,' New Zealand, "Application of international law."

138 Norway recognises that 'carrying out cyber operations with the intent of altering election results in another State, for example by manipulating election

systems... would be in violation of the prohibition of intervention', "Official Compendium […] A/76/136," 69.

139 Romania recognises that 'situations of tampering with the electoral processes in other States are relevant as a discussion under [the principle of non-intervention]', "Official Compendium […] A/76/136," 77.

140 The UK recognises 'the use by a hostile state of cyber operations to manipulate the electoral system to alter the results of an election in another state... Such acts must surely be a breach of the prohibition on intervention in the domestic affairs of states', Wright, "Cyber and International Law." The UK reaffirmed this position in 2021, 'the use of hostile cyber operations to manipulate the electoral system in another State to alter the results of an election... could... depending on the circumstances, be in violation of the international law prohibition on intervention', Foreign, Commonwealth & Development Office, "Application of international law."

141 The US considers 'a cyber operation by a State that interferes with another country's ability to hold an election or that manipulates another country's election results would be a clear violation of the rule of non-intervention', "Official Compendium […] A/76/136," 140.

142 Egan, "Remarks on international law."

143 Moynihan, "Application of international law to state cyberattacks," 41; Michael Schmitt, "Foreign cyber interference in elections: An international law primer, part I," *EJIL: Talk!* (blog), 16 October 2020; Tsagourias, "Electoral cyber interference" (2019).

144 Harriet Moynihan, "The application of international law to cyberspace: Sovereignty and non-intervention," *Just Security* (blog), 13 December 2019.

145 Puutio and Timis, "Deepfake Democracy;" Parkin, "The rise of the deepfake."

146 For example, see Scott Shane, "The fake Americans Russia created to influence the election," *The New York Times*, 7 September 2017, sec. U.S.

147 "Twitter, Facebook ban fake users; some had AI-created photos," *AP NEWS*, 21 December 2019.

148 Generally, see Cristian Vaccari and Andrew Chadwick, "Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news," *Social Media + Society* 6, no. 1 (January 2020).

149 Martijn Rasser, "Why are deepfakes so effective?" *Scientific American Blog Network* (blog), 14 August 2019.

150 Zakiah Koya, "Azmin's office insists videos are fake, refuting foreign reports," *The Star*, 17 June 2019.

151 See Samantha Bradshaw, Hannah Bailey, and Philip N. Howard, "Industrialized disinformation: 2020 Global inventory of organized social media manipulation" (Oxford Internet Institute, 2020).

152 Schmitt, "Autonomous cyber capabilities," 561.

153 Ibid., 561; also see Schmitt, *Tallinn Manual 2.0*, 318–319; Schmitt, "Grey Zones," 8.

154 New Zealand, "Application of international law."

155 "Official Compendium […] A/76/136," 69.

156 See Jamnejad and Wood, "The principle of non-intervention," 349.

157 See the predictions outlined in Philip Tully and Lee Foster, "Repurposing neural networks to generate synthetic media for information operations," *FireEye*, 5 August 2020.

158 Tsagourias, "Electoral cyber interference" (2020), 53–54.

159 Moynihan, "Application of international law to state cyberattacks," 41.

160 B.S. Murty, *Propaganda and World Public Order: The Legal Regulation of the Ideological Instrument of Coercion* (New Haven: Yale University Press, 1968), 29.

161 Watts, "Low-intensity cyber operations," 261 citing Murty; Murty, *Propaganda and World Public Order*, 129.
162 Watts, "Low-intensity cyber operations," 261–262.
163 Jamnejad and Wood, "The principle of non-intervention," 374.
164 Moynihan, "Application of international law to state cyberattacks," 41.
165 See previous discussion under *"interference" and "intervention"*.
166 As Brian Egan noted while serving as US Department of State Legal Adviser, 'a cyber operation by a State that interferes with another country's ability to hold an election or that manipulates another country's election results would be a clear violation of the rule of non-intervention', Egan, "Remarks on International Law;" Wright, "Cyber and International Law;" Australia, "2019 International Law Supplement;" the Netherlands in Staten-Generaal, "Letter of the Minister;" "International Law and Cyberspace, Finland's National Positions"; Germany, "On the Application."
167 'Blocking voting by cyber means, such as by disabling election machinery or by conducting a distributed denial of service attack, would likewise be coercive', Schmitt, ""Virtual" Disenfranchisement," 50; as '[i]n these situations, the result of the election, which is the expression of the freedom of choice of the electorate, is being manipulated against the will of the electorate', Michael N. Schmitt, "The use of cyber force and international law," in *The Oxford Handbook of the Use of Force in International Law* (Oxford University Press, 2015), 1116.
168 See Sean Watts, "International Law and Proposed U.S. Responses to the D.N.C. Hack," *Just Security* (blog), 14 October 2016; Duncan B. Hollis, "Russia and the DNC hack: What future for a duty of non-intervention?" *Opinio Juris* (blog), 25 July 2016; Ohlin, "Did Russian cyber-interference."
169 Philip Bump, "How Russian agents allegedly hacked the DNC and Clinton's campaign," *Washington Post*, 13 July 2018; Robert S. Mueller, "Report on the investigation into Russian interference in the 2016 presidential election," US Department of Justice, March 2019.
170 Bump, "How Russian Agents."
171 For example, see Delerue who states that '[t]he public release of the files aiming at coercing the United States in the conduct of the electoral process is likely to constitute a violation of the principle of non-intervention,' François Delerue, *Cyber Operations and International Law* (Cambridge University Press, 2020), 249–50; for Watts, 'if the network intrusion and extraction of information were conducted in order to assist a resistance or opposition movement's effort to influence political events in the target state, the intrusion would properly be considered for characterization as a violation of the principle of non-intervention', Watts, "*Low-Intensity Cyber Operations*," 256.
172 Jamnejad and Wood, "The principle of non-intervention," 374.
173 For a detailed breakdown of facts, see Mueller, "Report on the investigation."
174 Moynihan, "Application of International Law to State cyberattacks," 41; Jamnejad and Wood, "The principle of non-intervention," 374.
175 See Moynihan, "Application of international law to state cyberattacks," 42.
176 Ohlin, "Did Russian Cyber-Interference," 1593; Watts, "international law and proposed U.S. Responses;" Watts, "low-intensity cyber operations," 1580.
177 Steven J. Barela, "Cross-Border Cyber Ops to Erode legitimacy: An act of coercion," *Just Security* (blog), 12 January 2017; Steven J. Barela, "Zero Shades of Grey: Russian-ops violate international law," *Just Security* (blog), 29 March 2018 citing; Watts, "Low-Intensity Cyber Operations," 256.
178 Schmitt, "Grey zones," 8.
179 Ibid. (emphasis added).

180 The relevant part of the statement reads: 'Germany generally agrees with the opinion that malicious cyber activities targeting foreign elections may – either individually or as part of a wider campaign involving cyber and non-cyber-related tactics – constitute a wrongful intervention', 'cyber measures may constitute a prohibited intervention under international law if they are comparable in the scale and effect to coercion in the non-cyber contexts'…'it is conceivable that a State, by spreading disinformation via the internet, may deliberately incite violent political upheaval, riots and/or civil strife in a foreign country, thereby significantly impeding the orderly conduct of an election and the casting of ballots. Such activities may be comparable in scale and effect to the support of insurgents and may hence be akin to coercion in the above-mentioned sense', Germany "On the Application," 5.
181 See Schmitt, "Foreign Cyber Interference," (2020); and discussed in more detail in Schmitt, "Foreign Cyber Interference," (2021): 478–479.
182 Michael Schmitt, "Germany's Positions on International Law in Cyberspace Part I," *Just Security* (blog), 9 March 2021.
183 See Tully and Foster, "Repurposing Neural Networks."
184 Jamnejad and Wood, "The Principle of Non-Intervention," 381.
185 *Certain Activities*, 202.
186 See the positions of Australia, "Annex B;" Germany, "On the Application," 4–6; and New Zealand, "Application of International Law;" and the UK, Wright, "Cyber and International Law;" Foreign, Commonwealth & Development Office, "Application of International Law."
187 Including Australia, "Annex B;" Germany, "On the Application," 5; Israel in Schöndorf, "Israel's Perspective," 403; New Zealand, "Application of International Law;" Norway, in "Official Compendium […] A/76/136," 69; Romania, in "Official Compendium […] A/76/136," 77; the UK in Wright, "Cyber and International Law;" Foreign, Commonwealth & Development Office, "Application of International Law;" and the US in Ney, "DOD General Counsel;" "Official Compendium […] A/76/136," 140.

Bibliography

Australia's International Cyber Engagement Strategy. "2019 International Law Supplement to Australia's International Cyber Engagement Strategy, Annex A: Supplement to Australia's Position on the Application of International Law to State Conduct in Cyberspace." 2019. https://dfat.gov.au/international-relations/themes/cyber-affairs/aices/chapters/2019_international_law_supplement.html.

Australian Government. "Annex B: Australia's position on how international law applies to state conduct in cyberspace." *Australia's International Cyber and Critical Tech Engagement,* 2020. https://www.internationalcybertech.gov.au/our-work/annexes/annex-b.

"Appendix: International law in cyberspace, Letter of 5 July 2019 from the Minister of Foreign Affairs to the President of the House of Representatives on the international legal order in cyberspace, translation." *Minister of Foreign Affairs*, 5 July 2019. https://www.government.nl/binaries/government/documents/parliamentary-documents/2019/09/26/letter-to-the-parliament-on-the-international-legal-order-in-cyberspace/International+Law+in+the+Cyberdomain+-+Netherlands.pdf.

Foreign, Commonwealth & Development Office. "Application of international law to states' conduct in cyberspace." *UK Statement*, 3 June 2021. https://www.gov.uk/government/publications/application-of-international-law-to-states-conduct-in-cyberspace-uk-statement/application-of-international-law-to-states-conduct-in-cyberspace-uk-statement.

Barela, Steven J. "Cross-border cyber ops to erode legitimacy: An act of coercion." *Just Security* (blog), 12 January 2017. https://www.justsecurity.org/36212/cross-border-cyber-ops-erode-legitimacy-act-coercion/.

Barela, Steven J. "Zero shades of grey: Russian-ops violate international law." *Just Security* (blog), 29 March 2018. https://www.justsecurity.org/54340/shades-grey-russian-ops-violate-international-law/.

Bradshaw, Samantha, Hannah Bailey, and Philip N. Howard. "Industrialized Disinformation: 2020 Global inventory of organized social media manipulation." *Oxford Internet Institute*, 2020. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2021/02/CyberTroop-Report20-Draft9.pdf.

Buchan, Russell. *Cyber Espionage and International Law*. Oxford: Hart Publishing, 2019.

Bump, Philip. "How Russian agents allegedly hacked the DNC and Clinton's campaign." *Washington Post*, 13 July 2018. https://www.washingtonpost.com/news/politics/wp/2018/07/13/timeline-how-russian-agents-allegedly-hacked-the-dnc-and-clintons-campaign/.

Corn, Gary P., and Robert Taylor. "Sovereignty in the age of cyber." *AJIL Unbound* 111 (2017): 207–212. https://doi.org/10.1017/aju.2017.57.

Corten, Olivier, and Pierre Klein (eds). *The Vienna Conventions on the Law of Treaties—A Commentary*. Oxford University Press, 2011.

D'Amato, Anthony. "There is no norm of intervention or non-intervention in international law." *International Legal Theory* 7, no. 1 (2001): 9.

D'Amato, Anthony. "Trashing customary international law." *American Journal of International Law* 81, no. 1 (January 1987): 101. https://doi.org/10.2307/2202136.

Damrosch, Lori Fisler. "Politics across borders: Nonintervention and nonforcible influence over domestic affairs." *American Journal of International Law* 83, no. 1 (January 1989): 1. https://doi.org/10.2307/2202789.

Delerue, François. *Cyber Operations and International Law*. Cambridge University Press, 2020.

Egan, Brian. "Remarks on international law and stability in cyberspace." *Berkeley Law School, California*, 10 November 2016. https://2009-2017.state.gov/s/l/releases/remarks/264303.htm.

Fischerkepper, Michael P. "Current international law is not an adequate regime for cyberspace." *Lawfare* (blog), 22 April 2021. https://www.lawfareblog.com/current-international-law-not-adequate-regime-cyberspace.

Galston, William A. "Is seeing still believing? The deepfake challenge to truth in politics." *Brookings* (blog), 8 January 2020. https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/.

Gill, Terry D. "Non-intervention in the cyber context." In *Peacetime Regime for State Activities in Cyberspace*, edited by Katharina Ziolkowski. Tallinn: NATO CCD COE Publications, 2013.

"Hackers target Ukraine's election website." *Yahoo*, 25 October 2014. http://news.yahoo.com/hackers-target-ukraines-election-website-204128284.html.

Higgins, Rosalyn. "Intervention and international law." In *Intervention in World Politics*, edited by Hedley Bull. Oxford: Clarendon Press, 1984.

Higgins, Rosalyn. *Themes and Theories*. Oxford: Oxford University Press, 2009.

Hollis, Duncan B. "Russia and the DNC hack: What future for a duty of non-intervention?" *Opinio Juris* (blog), 25 July 2016. http://opiniojuris.org/2016/07/25/russia-and-the-dnc-hack-a-violation-of-the-duty-of-non-intervention/.

Hon. Paul C. Ney, Jr. "DOD General Counsel Remarks at U.S. Cyber Command Legal Conference." 2 March 2020. https://www.defense.gov/Newsroom/Speeches/Speech/Article/2099378/dod-general-counsel-remarks-at-us-cyber-command-legal-conference/.

"International Law and Cyberspace, Finland's National Positions." 2020. https://um.fi/documents/35732/0/KyberkannatPDF_EN.pdf/12bbbbde-623b-9f86-b254-07d5af3c6d85?t=1603097522727.

Jamnejad, Maziar, and Michael Wood. "The principle of non-intervention." *Leiden Journal of International Law* 22, no. 2 (June 2009): 345. https://doi.org/10.1017/S0922156509005858.

Jennings, R. Y., and Arthur Watts. *Oppenheim's International Law. Vol. 1: Peace*. 9th ed. London: Longman, 1996.

Joyner, Christopher C. "Coercion." In *Max Planck Encyclopedia of Public International Law*. Oxford University Press, December 2006.

Keller, Helen. "Friendly relations declaration (1970)." In *Max Planck Encyclopedia of Public International Law*. Oxford University Press, June 2009. https://doi.org/10.1093/law:epil/9780199231690/e938.

Kilovaty, Ido. "Doxfare: Politically motivated leaks and the future of the norm of non-intervention in the era of weaponized information." *Harvard National Security Journal* 9 (2018): 34.

Kilovaty, Ido. "The elephant in the room: Coercion." *AJIL Unbound* 113 (2019): 87–91. https://doi.org/10.1017/aju.2019.10.

Koya, Zakiah. "Azmin's office insists videos are fake, refuting foreign reports." *The Star*, 17 June 2019. https://www.thestar.com.my/news/nation/2019/06/17/azmin-office-insists-videos-are-fake-refuting-foreign-reports.

Kunig, Philip. "Intervention, prohibition of." *Max Planck Encyclopedia of Public International Law*, 16, April 2008.

Lowe, Vaughan. "The principle of non-intervention: use of force." In *The United Nations and the Principles of International Law: Essays in Memory of Michael Akehurst*, edited by Colin Warbrick. London: Routledge, 1994.

Lowe, Vaughan, and Antonios Tzanakopoulos. "Humanitarian intervention." In *Max Planck Encyclopedia of Public International Law*, Oxford University Press, May 2011.

Mak, Tim, and Dina Temple-Raston. "Where are the deepfakes in this presidential election?" *NPR*, 1 October 2020, sec. Investigations. https://www.npr.org/2020/10/01/918223033/where-are-the-deepfakes-in-this-presidential-election.

Metz, Rachel. "The fight to stay ahead of deepfake videos before the 2020 US election." *CNN*, 12 June 2019. https://www.cnn.com/2019/06/12/tech/deepfake-2020-detection/index.html.

Moulin, Thibault. "Reviving the principle of non-intervention in cyberspace: The path forward." *Journal of Conflict and Security Law* 25, no. 3 (1 December 2020): 423–447. https://doi.org/10.1093/jcsl/kraa011.

Moynihan, Harriet. "The application of international law to cyberspace: Sovereignty and non-intervention." *Just Security* (blog), 13 December 2019. https://www.justsecurity.org/67723/the-application-of-international-law-to-cyberspace-sovereignty-and-non-intervention/.

Moynihan, Harriet. "The application of international law to state cyberattacks: Sovereignty and non-intervention." Research Paper. *Chatham House, International Law Programme*, December 2019. https://www.chathamhouse.org/sites/default/files/2019-11-29-Intl-Law-Cyberattacks.pdf.

Mueller, Robert S. "Report on the investigation into Russian interference in the 2016 presidential election." *US Department of Justice*, March 2019.

Murty, B.S. *Propaganda and World Public Order: The Legal Regulation of the Ideological Instrument of Coercion*. New Haven: Yale University Press, 1968.

"Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by states," *UNODA, A/76/136*, August 2021. Accessed 10 February 2022. https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf.

Ohlin, Jens David. "Did Russian cyber-interference in the 2016 election violate international law?" *Texas Law Review* 95 (2017). https://doi.org/10.31228/osf.io/3vuzf.

"On the application of international law in cyberspace." *The Federal Government*, March 2021. https://www.auswaertiges-amt.de/blob/2446304/2ae17233b62966a4b7f16d50ca3c6802/on-the-application-of-international-law-in-cyberspace-data.pdf.

Parkin, Simon. "The rise of the deepfake and the threat to democracy." *The Guardian*, 22 June 2019. Accessed 27 June 2021. http://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy.

Puutio, Alexander, and David Alexandru Timis. "Deepfake democracy: Here's how modern elections could be decided by fake news." *World Economic Forum*, 5 October 2020. https://www.weforum.org/agenda/2020/10/deepfake-democracy-could-modern-elections-fall-prey-to-fiction/.

Rasser, Martijn. "Why are deepfakes so effective?" *Scientific American Blog Network* (blog), 14 August 2019. https://blogs.scientificamerican.com/observations/why-are-deepfakes-so-effective/.

Reisman, W.M. *Nullity and Revision: The Review and Enforcement of International Judgments and Awards*. New Haven: Yale University Press, 1971.

Rhodes, Samuel C. "Filter bubbles, echo chambers, and fake news: How social media conditions individuals to be less critical of political misinformation." *Political Communication* 39, no. 1 (2021): 1–22.

Rundle, James. "FBI warns deepfakes might become indistinguishable from reality." *Wall Street Journal*, 17 January 2020, sec. WSJ Pro. https://www.wsj.com/articles/fbi-warns-deepfakes-might-become-indistinguishable-from-reality-11579257004.

Schmitt, Michael N. "The law of cyber warfare: Quo vadis?" *Stanford Law & Policy Review* 25 (2014): 32.

Schmitt, Michael N. "The use of cyber force and international law." In *The Oxford Handbook of the Use of Force in International Law*, edited by Marc Weller, vol. 1. Oxford University Press, 2015. https://doi.org/10.1093/law/9780199673049.003.0053.

Schmitt, Michael N. "Grey zones in the international law of cyberspace." *Yale Journal of International Law* 42, no. 2 (2017): 1–21.

Schmitt, Michael N. "Peacetime cyber responses and wartime cyber operations under international law: An analytical vade mecum." *Harvard National Security Journal* 8 (2017): 44.

Schmitt, Michael. *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. 2nd ed., Cambridge: Cambridge University Press, 2017.

Schmitt, Michael N. ""Virtual" disenfranchisement: Cyber election meddling in the grey zones of international law." *Chicago Journal of International Law* 9 (2018): 39.

Schmitt, Michael N. "Taming the lawless void: Tracking the evolution of international law rules for cyberspace." *Texas National Security Review* 3, no. 3 (15 July 2020): 32–47.

Schmitt, Michael. "Foreign cyber interference in elections: An international law primer, part I." *EJIL: Talk!* (blog), 16 October 2020. https://www.ejiltalk.org/foreign-cyber-interference-in-elections-an-international-law-primer-part-i/.

Schmitt, Michael N. "Autonomous cyber capabilities and the international law of sovereignty and intervention." *International Law Studies* 96 (2020): 29.

Schmitt, Michael. "Germany's positions on international law in cyberspace part I." *Just Security* (blog), 9 March 2021. https://www.justsecurity.org/75242/germanys-positions-on-international-law-in-cyberspace/.

Schmitt, Michael N. "Foreign cyber interference in elections." *International Law Studies* 97 (2021): 27.

Schmitt, Michael. "Cybersecurity and international law." In *The Oxford Handbook of the International Law of Global Security*, edited by Robin Geiss and Nils Melzer, 661–678. Oxford University Press, 2021. https://doi.org/10.1093/law/9780198827276.003.0037.

Schmitt, Michael, and Jeffrey Biller. "Un-caging the bear? A case study in cyber opinio juris and unintended consequences." *EJIL Talk!* (blog), 24 October 2018. https://www.ejiltalk.org/un-caging-the-bear-a-case-study-in-cyber-opinio-juris-and-unintended-consequences/#more-16574.

Schmitt, Michael N, and Liis Vihul. "Respect for sovereignty in cyberspace." *Texas Law Review* 95 (2017): 1639.

Schmitt, Michael N., and Liis Vihul. "Sovereignty in cyberspace: Lex Lata Vel Non?" *AJIL Unbound* 111 (2017): 213–218. https://doi.org/10.1017/aju.2017.55.

Schöndorf, Roy. "Israel's perspective on key legal and practical issues concerning the application of international law to cyber operations." *International Law Studies* 97 (2021): 13.

Shane, Scott. "The fake Americans Russia created to influence the election." *The New York Times*, 7 September 2017, sec. U.S. https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election.html.

Simonite, Tom. "What happened to the deepfake threat to the election?" *Wired*, 16 November 2020. https://www.wired.com/story/what-happened-deepfake-threat-election/.

Staten-Generaal, Tweede Kamer der. "Letter of the Minister of Foreign Affairs to Parliament, 5 July 2019, Kamerstuk 33 694, Nr. 47, available (in Dutch)." Officiële publicatie, 5 July 2019. https://zoek.officielebekendmakingen.nl/kst-33694-47.html.

Tamkin, Emily. "Elections in Ghana marred by attempt to hack website and calls for the president to concede." *Foreign Policy* (blog), 8 December 2016. https://foreignpolicy.com/2016/12/08/elections-in-ghana-marred-by-hacked-website-and-calls-for-the-president-to-concede/.

New Zealand Ministry of Foreign Affairs and Trade. "The application of international law to state activity in cyberspace." 1 December 2020. https://www.mfat.govt.nz/en/media-and-resources/ministry-statements-and-speeches/cyber-il/.

Tsagourias, Nicholas. "Electoral cyber interference, self-determination and the principle of non-intervention in cyberspace." *EJIL Talk!* (blog), 26 August 2019. https://www.ejiltalk.org/electoral-cyber-interference-self-determination-and-the-principle-of-non-intervention-in-cyberspace/.

Tsagourias, Nicolas. "Electoral cyber interference, self-determination and the principle of non-intervention in cyberspace." In *Governing Cyberspace: Behavior, Power, and Diplomacy*, edited by Dennis Broeders and Bibi van den Berg, 45–64. Lanham: Rowman & Littlefield, 2020.

Tully, Philip, and Lee Foster. "Repurposing neural networks to generate synthetic media for information operations." *FireEye*, 5 August 2020. https://www.fireeye.com/blog/threat-research/2020/08/repurposing-neural-networks-to-generate-synthetic-media-for-information-operations.html.

"Twitter, Facebook ban fake users; some had AI-created photos." *AP NEWS*, 21 December 2019. Accessed 16 June 2021. https://apnews.com/article/ap-top-news-elections-artificial-intelligence-social-platforms-us-news-7c9fd-798212cac63925d205142e811ea.

Tzanakopoulos, Antonios. "The right to be free from economic coercion." *Cambridge Journal of International and Comparative Law* 4, no. 3 (2015): 616–633. https://doi.org/10.7574/cjicl.04.03.616.

Vaccari, Cristian, and Andrew Chadwick. "Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news." *Social Media + Society* 6, no. 1 (January 2020): 205630512090340. https://doi.org/10.1177/2056305120903408.

Watts, Sean. "International law and proposed U.S. responses to the D.N.C. Hack." *Just Security* (blog), 14 October 2016. https://www.justsecurity.org/33558/international-law-proposed-u-s-responses-d-n-c-hack/.

Watts, Sean. "Low-intensity cyber operations and the principle of non-intervention." In *Cyber War: Law and Ethics for Virtual Conflicts*, edited by Jens David Ohlin, Kevin Govern, and Claire Finkelstein, Oxford: Oxford University Press, 2019.

Wright, Attorney General Jeremy. "Cyber and international law in the 21st century," *Chatham House*, 23 May 2018. https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century.

Wright, Quincy. "Subversive intervention." *American Journal of International Law* 54, no. 3 (1960): 521–535.

Ziegler, Katja S. "Domaine Réservé." In *Max Planck Encyclopedia of Public International Law*. Oxford University Press, April 2013.

Ziolkowski, Katharina. "General principles of international law as applicable in cyberspace." In *Peacetime Regime for State Activities in Cyberspace: International Law, International Relations and Diplomacy*, edited by Katharina Ziolkowski. Tallinn, Estonia: NATO CCD COE Publication, 2013.

Index

Note: *Italic* page numbers refer to figures and page numbers followed by "n" denote endnotes.