NATO@70

NDC Research Paper

Research Division – NATO Defense College – Rome

The Brain and the Processor:
Unpacking the Challenges of Human-Machine Interaction

Edited by
Andrea Gilli

Research Division

No. 06
Dec 2019

The Brain and the Processor: Unpacking the Challenges of Human-Machine Interaction

Edited by
Andrea Gilli

NATO DEFENSE COLLEGE

NATO Defense College Cataloguing in Publication-Data:
The Brain and the Processor: Unpacking the Challenges of Human-Machine Interaction
(NATO Defense College "*NDC Research Papers Series*")

NDC Research Paper 6
Edited by Andrea Gilli
Copy-editing: Mary Di Martino
Series editor: Thierry Tardy

ISSN: 2618-0057
ISSN (online): 2618-0251

NDC 2019

Limited copies of this NDC Research Paper are available and may be obtained directly from
NATO Defense College, Research Division
Via Giorgio Pelosi, 1 - 00143 Rome, Italy
Fax +39-06-50 52 57 97
E-mail: publications@ndc.nato.int
Website: http://www.ndc.nato.int
Follow us on twitter: https://twitter.com/NDC_Research

Printed and bound by
http://www.lightskyconsulting.com/

Table of contents

Contributors

Ajay AGRAWAL is the Geoffrey Taber Professor of Entrepreneurship and Innovation at the University of Toronto's Rotman School of Management, Research Associate at the National Bureau of Economic Research (NBER), and founder of the Creative Destruction Lab.
Professor Agrawal conducts research on entrepreneurial finance, the economics of science, the commercialization of university science, and the economic geography of innovation, recently turning his attention to the economics of Artificial Intelligence (AI). In addition to a number of scholarly articles on the topic, he has co-authored *Prediction machines: the simple economics of Artificial Intelligence* (Harvard Business School Press, 2018) as well as co-edited *The economics of Artificial Intelligence: an agenda* (University of Chicago Press, 2019). Professor Agrawal serves on the editorial board of the *Strategic Management Journal, Management Science,* and the *Journal of Urban Economics.*

Aaron W. CELAYA is an Air Force Institute of Technology Doctoral Student, New College, Department of Experimental Psychology, University of Oxford, United Kingdom. He researches metacognition through decision confidence in human-machine teaming, with an emphasis on algorithmic advisement sources. This research is conducted within the Attention and Cognitive Control Laboratory which is part of the Department of Experimental Psychology at the University of Oxford. Prior to his current assignment, he was the Research Center Director of Operations and Senior Instructor, Warfighter Effectiveness Research Center, Department of Behavioral Sciences and Leadership, Dean of Faculty, United States Air Force Academy, Colorado.

Steven FINO is a Colonel and command pilot in the US Air Force. A distinguished graduate of the US Air Force Academy, he earned advanced degrees from the University of California-Los Angeles, Air Force Institute of Technology, and Air Force School of Advanced Air and Space Studies before earning his doctorate from the Massachusetts Institute of Technology. He has ten years' experience flying the F-15C Eagle, including two years as an operational test pilot, and is a graduate of the USAF Weapons School. Additionally, he has two years' experience piloting the RQ-4B Global Hawk and served as the commander for a combined U-2 and RQ-4B unit. He has served in the Pentagon as an Airpower Strategist for the Chief of Staff of the USAF, an operations analyst for the Secretary of Defense, and a Division Director for the Under Secretary of Defense for Intelligence. He is the author of *Tiger check: automating the US Air Force Fighter Pilot in air-to-air Combat, 1950-1980* (Johns Hopkins University Press, 2017).

Joshua GANS is a Professor of Strategic Management and holder of the Jeffrey S. Skoll Chair of Technical Innovation and Entrepreneurship at the Rotman School of Management, University of Toronto (with a cross appointment in the Department of Economics). He is also Chief Economist of the University of Toronto's Creative Destruction Lab. Joshua holds a Ph.D. from Stanford University and an honors degree in economics from the University of Queensland. In 2012, Joshua was appointed as a Research Associate of the National Bureau of Economic Research (NBER) in the Productivity, Innovation and Entrepreneurship Program. At Rotman, he teaches MBA students entrepreneurial strategy. He has also co-authored (with Stephen King and Robin Stonecash) the Australasian edition of Greg Mankiw's *Principles of economics* (published by Cengage), *Core economics for managers* (Cengage), *Finishing the job* (MUP), *Parentonomics* (New South/MIT Press), *Information wants to be shared* (Harvard Business Review Press), *The disruption dilemma* (MIT Press, 2016); *Scholarly publishing and its discontents* (2017), and *Prediction machines: the simple economics of Artificial Intelligence* (HBR Press, 2018). His most recent book is *Innovation + equality: creating a future that is more Star Trek than Terminator* (MIT Press, 2019).

Andrea GILLI is a Senior Researcher at the NATO Defense College (NDC) in Rome where he works on issues related to technological change and military innovation. He holds a PhD in Social and Political Science from the European University Institute (EUI) in Fiesole (Florence, Italy), an MSc in International Relations from the London School of Economics and Political Science (LSE) and a BA in Political Science and Economics from the University of Turin (Italy). In the past, Andrea held visiting and post-doctoral fellowships at Johns Hopkins University as well as Columbia, Stanford and Harvard University. Andrea also worked for the European Union Institute for Security Studies in Paris and the US Department of Defense's Office for Net Assessment. His research has been published or is forthcoming in *International Security*, *Security Studies* and *Journal of Strategic Studies*.

Avi GOLDFARB is the Rotman Chair in Artificial Intelligence and Healthcare, and Professor of Marketing, at the Rotman School of Management, University of Toronto. Avi is also Chief Data Scientist at the Creative Destruction Lab, Senior Editor at Marketing Science, and a Research Associate at the National Bureau of Economic Research. Avi received his PhD in economics from Northwestern University. His research focuses on the opportunities and challenges of the digital economy. He has published over 80 academic articles in a variety of outlets in marketing, statistics, law, management, medicine, computing, and economics. This research has been discussed in policy documents from the US Congress, the White House, the European Commission, and other government organizations, as well as in the press including the *New York Times* and the *Economist*. Along with Ajay Agrawal and Joshua Gans, Avi is the author of the book *Prediction machines: the simple economics of Artificial Intelligence* (HBR Press, 2018).

Steven HILL is NATO Secretary General Jens Stoltenberg's chief legal counsel. He leads a multinational team that provides legal advice on the whole range of issues facing the 29-member Alliance. He participates in NATO's current work on cyber issues, including as a member of NATO's Cyber Defence Management Board. He was NATO's observer in the International Group of Experts for the Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations. He regularly speaks to legal audiences and takes part in a range of expert working groups on cyber law issues.
Prior to joining NATO in February 2014, Steve was Counselor for Legal Affairs at the United States Mission to the United Nations. He has served as Professor of Law at the Hopkins-Nanjing Center in China (2010-2011), Professional Lecturer in Law at SAIS (2007), and Adjunct Professor of Law at George Mason University School of Law (2002-2004).
Steve graduated from Yale Law School and Harvard College and is a member of the New York bar, the Executive Council of the American Society of International Law, and the *Société française pour le droit international*. He serves on the boards of the International Society of Military Law and Law of War and the Center for International Business and Human Rights has been a term member of the Council on Foreign Relations.

Richard KELLY is a Technology Ethics Specialist within PricewaterhouseCoopers (PwC) Australia's Emerging Technology Team. He has a key focus on technology ethics and responsible Artificial Intelligence with more than 15 years of experience developing and implementing small to large Information and Communication Technology (ICT) projects across a variety of industries. Richard was instrumental in developing the ethical AI framework and governance approach for PwC's Global Responsible AI offering.

Nadia MARSAN has been a member of NATO's International Staff since 2006, joining the NATO Office of Legal Affairs in August 2017. Prior to joining NATO, Mrs. Marsan held various legal adviser positions, including in electoral complaints adjudication and in refugee status determination. Mrs. Marsan holds a BA in Philosophy from McGill University and an LLB/JD from the University of Ottawa, Faculty of Law.

Antonio MISSIROLI is NATO's Assistant Secretary General for Emerging Security Challenges. Prior to joining NATO, he was the Director of the European Union Institute for Security Studies (EUISS) in Paris (2012-17). Previously, he was Adviser at the Bureau of European Policy Advisers (BEPA) of the European Commission (2010-2012); Director of Studies at the European Policy Centre in Brussels (2005-2010), and Senior Research Fellow at the W/EU Institute for Security Studies in Paris (1998-2005). He was also Head of European Studies at CeSPI in Rome (1994-97) and a Visiting Fellow at St Antony's College, Oxford (1996-97). As well as being a professional journalist, Dr Missiroli has also taught at Bath and Trento as well as Boston University, SAIS/Johns Hopkins, at the College of Europe (Bruges) and Sciences Po (Paris). Dr. Missiroli holds a PhD in Contemporary History from the Scuola Normale Superiore (Pisa) and a Master's degree in International Public Policy from SAIS/Johns Hopkins University.

Massimo PELLEGRINO is a partner at PricewaterhouseCoopers (PwC), Italy, where he is in charge of New Ventures, the business unit focused on innovation and emerging technologies. He is also the Global Leader of the Ethical AI and Governance practice. Prior to joining PwC, Massimo was the Global Vice President of the Enterprise Information Solutions business unit at Hewlett Packard. His areas of competence are Artificial Intelligence, Advanced Analytics and Innovation Management.

Catherine WARNER is Director, Centre for Maritime Research and Experimentation, an executive body of NATO's Science and Technology Organization, since 2017. Dr Warner is a Senior Executive and Technical Leader with over 25 years' experience in defence science and technology, in particular in studies, analyses, and operational test and evaluation. Previously, Dr Warner was the Science Advisor to the Director, Operational Test and Evaluation within the US Office of the Secretary of Defence where she was the primary technical advisor to the Director, military, and civilian staff. Dr Warner deployed to Afghanistan during 2013-2014 in support of the International Security Assistance Force (ISAF), as the Senior Telecommunications Advisor to the Afghan government. Her portfolio included spectrum management, cybersecurity, and optical fiber critical infrastructure. Prior to joining the Defence Department, Dr Warner worked at the not-for-profit Institute for Defense Analyses (IDA) as a research analyst, project leader, and assistant director in the Operational Evaluation Division. Dr Warner is originally from Albuquerque, New Mexico, and she earned a PhD in Chemistry from Princeton University.

Nick YEUNG is professor of cognitive neuroscience in the Department of Experimental Psychology at the University of Oxford, where he heads the Attention and Cognitive Control Laboratory. He received his PhD from the University of Cambridge in 2000, and has held positions at Princeton University and Carnegie Mellon University. His research investigates human attention, memory, and decision making using a combination of behavioural, computational, and brain imaging techniques. A current focus of his lab's research is on how individuals seek and share information when making decisions in groups and teams, and how trust develops among team members. A key application of this research is to understand and improve human-machine teaming. Prof. Yeung has published more than 70 articles, based on research funding that has included projects with Defence Advanced Research Project Agency (DARPA), Intelligence Advanced Research Project Agency (IARPA) and the UK Defence Science and Technology Laboratory.

List of abbreviations

ACT	Allied Command Transformation
AFRA	Air Force Research Intelligence
AI	Artificial Intelligence
BD	Big Data
BEPA	Bureau of European Policy Advisers
CSAF	Chief of Staff US Air Force
DARPA	Defense Advanced Research Projects Agency
DoD	Department of Defense
EU	European Union
EUISS	European Union Institute for Security Studies
GCS	Ground Control Station
GIB	Guy-in-Black
HOTAS	Hands-on-Throttles-And-Stick
IARPA	Intelligence Advanced Research Projects Agency
IBM	International Business Machines Corporation
ISR	Intelligence, Surveillance, Reconnaissance
LAWS	Lethal Autonomous Weapon System
MIT	Massachussets Institute of Technology
MSFRIC	Muir S. Fairchild Research Information Center
NATO	North Atlantic Treaty Organization
NBER	National Bureau of Economic Research
NCO	Non-commissioned officer
NDC	NATO Defense College
NGO	Non-Governmental Organization
OECD	Organisation for Economic Co-operation and Development
OODA	Observe-Orient-Decide-Act
PwC	PricewaterhouseCoopers
RPA	Remotely Piloted Aircraft
STEM	Science, Technology, Engineering and Maths
UAV	Unmanned Aerial Vehicle
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization
US	United States
USAF	United States Air Force
USD	US Dollars

Acknowledgements

This volume was made possible thanks to a workshop jointly organized by the NDC and Paris-based IRSEM *(Institut de Recherche Stratégique de l'École Militaire)*, on 9-10 December 2018 in Rome, on "The future of warfare".

It is this workshop that inspired the volume, brought to fruition in the intervening months by the dedication and cooperation of many people.

As project leader, I would like first and foremost to thank the contributors to this volume, for their vision and support, and for their commitment and patience during the months of revision and the final editing process.

A special note of thanks is extended to all the other conference participants for their insights and comments: Ronald Arkin, Sergio Barretta, Nehal Bhuta, Edward Christie, Fabio Corona, Federica Dall'Arche, Anja Dahlmann, François Delerue, Ruben Diaz-Playa, Kara Frederik, Robin Geiss, Andrea Giuricin, Jean-Baptiste Jeangene Vilmer, Jan Mazal and Martina Toplicanec.

I would also like to thank the several anonymous reviewers who provided comments and feedback on different chapters, as well as the many people I spoke with on this and related topics at the NATO Defense College, Bocconi University, ETH-Zurich, the US Department of Defense's Joint Artificial Intelligence Center, Georgetown's Center for Security and Emerging Technology, the Center for Naval Analysis, NATO HQ and several other friends and colleagues who were a constant source of support and inspiration. In particular, special thanks to Thierry Tardy, Mary Di Martino, Elsa Kania, Mauro Gilli, Eugenio Mengarini, Isabella Silvestri, Nina Silove, Ann-Sophie Leonard, Lisa Becker, Zoe Stanley-Lockman, and Sophie-Charlotte Fisher, for the constant support throughout the project.

Andrea Gilli
Senior Researcher
NATO Defense College

Foreword

The 21st century has seen an unprecedented growth and acceleration of technological development, thanks in large part to the commercial sector, and especially in the digital domain. Shrewd use of readily available new technologies, often in combination, has offered state and non-state actors a multitude of different ways to inflict damage and disruption well above and beyond what was imaginable a few decades ago. Such damage and disruption can now be inflicted not only on traditionally superior military forces on the battlefield, but also on the civilian population and critical infrastructure, for example by cyberattacks. In addition, most of these technologies are developed at a far remove from the traditional defence industry pipeline, and are characterised by an extremely short time to market.

Emerging and disruptive technologies – starting with data analytics, artificial intelligence and autonomy – will change the way we (or potential adversaries) operate. Yet it is still unclear exactly when and how this will happen, and with what implications. Futur(olog)ists often underline this point: Roy Amara coined the famous "law" whereby we tend to overestimate the effect of a new technology in the short term and to underestimate it in the long term, while Alvin Toffler quipped that "the future always comes too fast, and in the wrong order". What is sure is that these technologies – which have to date enjoyed only limited application in the military domain – also have the potential to change balances of power, among countries as well as other actors. Access to technologies and clear ideas about their intended use will also play a key role, raising important questions about values and norms.

While generating new risks and threats to our security, however, these technologies also bring significant opportunities in many different fields. It is therefore vital that we understand them, analyse what implications they have for our security, decide how we want to take advantage of the opportunities they offer, and ensure that we are ready to mitigate the resulting risks. Over the past 70 years, NATO's technological edge has been an essential enabler of its military dominance. Developments in emerging and disruptive technologies are driven by national governments and, increasingly, by the private sector. At the same time, they have a profound impact on the ability of the Alliance to carry out its core tasks. This is related in particular to their impact on a) deterrence and defence (they can change the nature of military conflict and the way we conduct military operations); b) interoperability (risk of a technological gap between Allies); c) mobility and communications (vulnerability of critical infrastructure); and d) arms control mechanisms (verification would prove difficult).

NATO's Defence Ministers held an informal discussion on these issues in June 2019, with a view to agreeing a dedicated roadmap. Similarly, the EU Defence Ministers (22 EU members are also NATO members) debated emerging and disruptive technologies in August of the same year. There is little doubt that their application will represent a major policy challenge for everybody in the years to come. Especially among diplomats and political elites, there is an innate and understandable fear of loss of control – as exemplified by the ongoing controversy over Lethal Autonomous Weapon Systems (LAWS) or "killer robots" – and a sense of powerlessness in the face of new and often little understood new technologies. One need only recall the shock generated by chess champion Garry Kasparov's defeat by IBM's Deep Blue computer in 1997 or, 20 years later, by Google Deep Mind's victory over two of the best players of the ancient and fiendishly complex Chinese game of Go. It is indeed difficult to dismiss Henry Kissinger's concerns about the "end of Enlightenment" – when machines substitute our brains and not just our muscles – as well as the warnings raised by the likes of Stephen Hawking or even Elon Musk. Still, it is equally important to engage in a rational discussion on the potential advantages that they can already bring, as is already the case for instance in healthcare. In fact, we already make use of semi-autonomous intelligence in the military domain for primarily defensive purposes, e.g. with naval vessels using sensors to detect and intercept attacking devices. Similarly, in the cyber domain, AI is already widely used for detection – in order quickly to discern between normal and anomalous online behavior – and prediction (swift identification of patterns). By the same token, law enforcement uses sophisticated algorithms to process large sets of data and track down suspicious activities.

Prediction, detection and disablement are thus functions where AI-related technologies and applications are already worth considering and developing. Their predominantly defensive and tactical nature – encompassing situational awareness, anticipation and mitigation – should not, in principle, raise insurmountable ethical dilemmas or legal concerns – although these challenges will not be easy or immediate to solve. AI allows systems to learn of, and from, cyberattacks faster. It can also improve the realism of crisis management exercises, forcing our defenders and first responders to operate in a more challenging environment, honing their skills. And AI can facilitate both force protection and civil protection in theatres, especially densely populated areas. Even in terms of decision-making, AI can help overcome all too human shortcomings like groupthink, confirmation bias, bureaucratic politics and poor risk judgement.

Having said this, two main challenges remain, for both the Alliance and individual Allies. In terms of decision-making, especially at strategic level, our multi-stakeholder and consensual practices may clash with the "fast and furious" environment shaped by the new technologies. In terms of procurement and acquisitions, the current and foreseeable speed of technological R&D substantially defies the lengthy procedures that still characterize our public policies and risks creating a structural weakness.

This volume is particularly helpful in the first respect. Its different chapters guide the reader through the challenges that human beings, leaders, organizations and communities in general face when interacting with new technologies: what does it mean for new organizations to adopt new technologies, both in terms of division of labour between machines and human beings and in terms of organizational structure? Similarly, how do we react to autonomous machines and, more importantly, is our brain prone to errors or mistakes? Does technological change affect group identities and, consequently, organizational performance? The first three chapters provide food for thought and some preliminary answers to the questions. The last two go into greater depth on the ethical and legal debates highlighting the implications of automation in these two domains.

Antonio Missiroli
NATO's Assistant Secretary General
for Emerging Security Challenges

Introduction

Andrea Gilli

The age of intelligent machines opens up enormous opportunities but also generates some risks. The bulk of the media and scholarly analyses, especially when it comes to security issues, are more prone to emphasize the dangers, unfortunately. Skepticism and pessimism thus dominate the debate, and consequently most observers take for granted that this wave of technological transformation will lead to growing instability, inadvertent escalation, and more intense conflicts. Little space is often left for more nuanced and balanced assessments.[1] This is unfortunate because, at closer scrutiny, some, if not most, of these worries appear highly exaggerated.[2] This does not mean, however, that the future is bereft of challenges. As the recent book *The Technology Trap* by Oxford Prof. Carl Benedikt Frey shows, across history, human beings have reacted differently to technological change, depending whether it is more enabling or labor-replacing.[3] Labor-replacing technologies have generated opposition, ultimately slowing down the course of human progress in a dramatic way. Enabling technologies, in contrast, have been more easily embraced and accepted. When we look toward the future, this simple insight has important implications: should artificial intelligence, machine learning and big data prove to be labor-replacing technologies, the political opposition they will generate, rather than their intrinsic properties, will likely determine the future distribution of power and wealth around the world.

This volume touches upon this and a set of related messages that ultimately put human beings at the very center of this transformation toward intelligent machines: one of the paradoxes of the second machine age – as Erik Brynjolfsson and Andrew McAfee call it – is in fact that the more tasks computers take over, the more important becomes the role of people.[4] Technology replaces tasks, not human beings, and human beings at each

1 One of the paradoxes is thus that social scientists are more interested in so-called lethal autonomous weapons than military officers, one reason being that the latter have a clear sense of the political risks related to their employment as well as of their limited tactical effectiveness in many circumstances.

2 A. Gilli, "Preparing for 'NATO-mation': the Atlantic Alliance toward the age of Artificial Intelligence", *NDC Policy Brief,* No.4, February 2019.

3 C. B. Frey, *The technology trap: capital, labor, and power in the age of automation,* Princeton, Princeton University Press, 2019.

4 E. Brynjolfsson and A. McAfee, *The second machine age: work, progress, and prosperity in a time of brilliant technologies*, New York, W. W. Norton & Company, 2014.

round of automation specialize in the tasks with higher added value.[5] In other words, technological changes tend to complement human activities, enabling people to specialise in those areas where they have a comparative advantage. However, the transition is not bereft of problems: new skills and new organizational structures are needed to exploit the potentials of new technologies.

The emergence of artificial intelligence, machine learning and big data – that in this volume will be interchangeably used alongside terms such as intelligent machines – can potentially represent a major turning point in human history. For the moment, we are dealing – and we will likely continue to do so – with weak or limited AI, namely single-purpose systems, i.e. machines that may reach an outstanding performance in one realm but cannot even fulfill the most basic task in a related field.[6] Strong, or general, artificial intelligence – also referred as singularity – will occur when machines will, like human beings, be able to conduct a plurality of multidimensional activities, from driving a car to empatizing with peers, from playing video games to conducting different work activities.[7]

The contributions in this volume attempt to map the underlying causes, direct implications and broader consequences of the transformation at hand. Their main collective value is that they adopt different, but complementary approaches that ultimately corroborate the underlying theme of the entire volume, namely that as machines become more capable and easily available, the more important become human beings. The world of defense and security often looks with both amazement and concern at other fields. It rarely, however, truly engages with them, also intellectually.

Building on a conference jointly organised in December 2018 between the NATO Defense College Research Division and the Paris-based *Institut de Recherche Stratégique de l'Ecole Militaire*, this volume has tried to address this shortfall by bringing together defense scholars and security practioners, economists and psychologists, historians and lawyers as well as business consultants and roboticists with the goal of looking from different angles at the domain of AI and reason, together, on its possible meanings and consequences.

In the first chapter, Ajay Agrawal, Joshua Gans and Avi Goldfarb elaborate on the intuition that machines complement the work of human beings. Machine learning is

5 They can also specialize on those whose value-added is so limited that there is not a business case for automation. This is an important issue for social policy but it will not be discussed in this volume.

6 A. Asoni *et al.*, "A mercenary army of the poor? Technological change and the demographic composition of the post-9/11 US Military", *Journal of Strategic Studies*, 2019 (forthcoming).

7 *Ibid.*

prediction – they say. Technological progress makes prediction cheaper.[8] As a result, demand for it increases along with its complements like data, action and judgement. Complements are those goods and services necessary to exploit in full all the benefits of a certain product.[9] However, while access to prediction technologies may be more or less even, not all actors are equally endowed to meet the increased demand in complements. This disparity will affect competitive advantages in the years ahead. Agrawal, Gans and Goldfarb are leading economists who, mostly, focus on companies and markets but their message applies equally to the political world, and in particular to the field of security and defense.[10] Developing, protecting, feeding and nurturing the complements that machine learning requires will become more and more important in the future as intelligent machines are adopted more pervasively. This message is important for both NATO and its member states. The chapter by Agrawal, Gans and Goldfarb also has a broader message, however: technology cannot be a substitute for strategy. This is, interestingly, a lesson we can draw from the first machine age, the era of the steam engine when France systematically tried, and failed, to defeat Great Britain through new (naval) technologies rather than by developing more appropriate strategies.[11] The chapter leaves open the question of who can win a strategic competition in the age of intelligent machines, and how. This should be the focus of current and future generation (grand) strategists.

Using and interacting with intelligent machines, however, is not straightforward. Not all machines are created equal, and not all human beings are equally capable to deal with machines. Nicholas Yeung and Aaron Celaya summarise, in their chapter, the existing literature in psychology, neuroscience and related disciplines on the human-machines interaction to highlight the challenges looming ahead in integrating artificial intelligence. Human beings, as Nobel-laureate Daniel Kahneman has shown long ago, do not necessarily act rationally and consistently.[12] This is a key factor to take into consideration when machines work next to human beings – and this will be the case so long as we do not move from weak to strong AI. First, machines must be designed on the basis of operators' perceptions and

8 Their chapter is based on their book A. Agrawal *et al.*, *Prediction machines: the simple economics of Artificial Intelligence,* Cambridge, Harvard Business School Press, 2018.

9 For a discussion of the role of complements in the world of defense and security, see A. Gilli and M. Gilli, "The Diffusion of drone warfare? Industrial, infrastructural and organizational challenges: military innovations and the ecosystem challenge", *Security Studies*, Vol.25, No.1, 2016, pp.50-84.

10 An additional focus on public policy is presented in A. Agrawal *et al.*, *The economics of Artificial Intelligence: an agenda*, Chicago, University of Chicago Press, 2019.

11 R. Gardiner and A. Lambert, *Steam, steel and shellfire: the steam warship, 1815-1905*, London, Conway's History, 2001.

12 D. Kahneman, *Thinking, fast and slow,* New York, Farrar, Straus and Giroux, 2011.

features – not solely on the basis of what engineers aim to achieve. We can have perfect machines, but they will prove useless if they are not in sync with the way our brains work. Second, and related, if intended to complement the work of human beings, machines will have to address, not reinforce the biases and flaws characterising human behaviour. Unfortunately, recent stories about AI show that this is not easy at all. When it comes to combat, this may represent a huge challenge, especially as adversaries employ cover, concealment and camouflage tactics to neutralize their adversaries' capabilities.[13] Yeung and Celaya's chapter has a third merit: more or less implicitly, it suggests that education, recruitment and training will have to adapt as we integrate more intelligent machines into our processes. This brings back to some of the implications of Agrawal, Gans and Goldfarb's work about organisational adaptation. Also in this case, these considerations have wide security implications as these challenges will equally affect NATO and its Allies' Ministries of Defense. More prosaically, as artificial intelligence advances, we need a better understanding of human intelligence to ensure that we get all the benefits from this wave of technological change.

Artificial intelligence, machine learning and big data necessarily push us to think toward the future. Is there, however, any past instance that can help us in this endeavour? The chapter by Steven Fino tells the history of automation in air-to-air warfare during the Cold War. The chapter has primarily a historical, anthropological and science and technology studies perspective.[14] However, it directly speaks to the two previous chapters as well. How did US fighter pilots react to automation in air-to-air warfare? There were, as Agrawal, Gans and Goldfarb note in their chapter, complementary challenges. As combat aircraft became more sophisticated, and more tasks were automated, selection of fighter pilots became stricter and the skills required more advanced. For instance, nowadays, it can take up to seven years to train a pilot (already selected through several mechanisms), making pilots at times a more scarce resource than the expensive weapon systems they sit on.[15] There are also cognitive, psychological and by extension anthropological reactions that Fino documents and that were triggered by change in technology and thus in tasks. A recurrent dilemma in the Air Force concerns, for instance, the tension between flying an airplane and fighting with an airplane. Technological progress altered the balance between these tasks,

13 S. D. Biddle, *Military power: explaining victory and defeat in modern warfare*, Princeton, Princeton University Press, 2004.

14 For a broader and deeper perspective, see S. A. Fino, *Tiger check: automating the US Air Force Fighter Pilot in air-to-air combat, 1950-1980*, Baltimore, Johns Hopkins University Press, 2017.

15 A. Asoni *et al.*, "A mercenary army of the poor? Technological change and the demographic composition of the post-9/11 US Military".

thus allowing bureaucracy and politics to step in: pilots opposed the automation of combat, as this deprived them of a key component of their identity, for example. This is the political opposition mentioned at the beginning in speaking about Frey's monograph but it is also related to Celaya and Yeung's contribution on psychology. Importantly, although Fino's work looks into the past, the implications of his work directly point toward the future as, realistically, we will see the same challenges and dynamics in the years ahead. The chapter by Fino, however, provides also another important contribution that goes beyond this volume. It shows that to understand the future of tomorrow we can gain several lessons from the future of yesterday. We live, however, in an age in which history, and especially business, diplomatic and military history are increasingly going out of fashion. This is a serious problem that Fino's chapter should encourage us further to address, scholars and policy-makers alike.[16]

The chapter by Andrea Gilli, Massimo Pellegrino and Richard Kelly highlights one of the biggest paradoxes of our current age of intelligent machines. As machines take over more and more tasks, human beings need to focus on ethics to enable them to work as expected – not only in terms of efficiency and effectiveness but also in terms of legality, ethics and morality. Choices and decisions are based on a plurality of factors: ethics is among the most important. As machines increasingly take decisions, a set of important ethical questions beg for an answer. This chapter aims to provide a broad introduction to the issue of ethics and AI. Probably the most surprising part of this contribution is how practical these questions and challenges are – although raised on the basis of abstract principles. There is a related consideration that brings back to Chapter 1 from Agrawal, Gans and Goldfarb. Data is a key input. However, having the authorisation to gather and store, as well as ensuring that such data are not biased are not only ethical imperatives but also key conditions to turn data into an asset, whether in the military or economic field.

Ethics helps us define the purpose of our life as well as the driving principles of our action. Ethics, in this sense, also helps us define the parameters of law. The final chapter by Steven Hill and Nadia Marsan is an important contribution in this respect. Human relations are complex, especially when they occur on very unstable ground. Rapid technological change in the context of the use of force within and among sovereign states is a particularly delicate issue: but it is also one of the most important topics of our day and age. For a significant, and growing, part of humanity, life is no longer short, nasty and brutish.

16 H. Brands and F. J. Gavin, "The historical profession is committing slow-motion suicide", *War on the Rocks,* 10 December 2018, https://warontherocks.com/2018/12/the-historical-profession-is-committing-slow-motion-suicide/ (accessed 6 November 2019).

Technological change is definitely responsible for these achievements. But the world we live in is not only made of material interactions. The social world also plays a prominent role. International law is a key component of this social world, especially in the international arena as Marsan and Hill remind us by delving into the implications of artificial intelligence for international law and, in particular, the laws of war. Their perspective highlights the role that international law, and its community, can play to address different challenges that new technologies can generate. International law is not, by definition, the only arbiter but it can help reach common understandings among the key stakeholders. In fact, by providing a common background, international law can help find common, or at least coordinated, solutions. This is where the role of NATO can be particularly important as its Member States, through the Alliance, can promote some norms and achieve the ratification of some laws for the global goal of international security. There are several challenges ahead, however. Some are objective and refer to the inherent difficulties stemming from this field. Others are more subjective, even intersubjective and relates to the way actors come to understand the world we live in. Marsan and Hill start from the concept of legal interoperability to discuss how the principle of the rule of law is at the heart of NATO operations and thus shapes our interpretation of different applications of artificial intelligence – from intelligence, surveillance and reconnaissance (ISR) to disinformation campaigns.

The concluding chapter is left to Catherine Warner who has accepted the challenging task of building on these important contributions, connect their different messages and elaborate what this means for international security, in general, and NATO in particular. Her concluding remarks emphasize, among others, the importance of research and, more broadly, of science and technology – if NATO wants to remain not only capable of dissuading, deterring and defending in the years ahead but also of shaping to its benefits these very technological, economic and social developments.[17] In other words, Warner reminds us what Abrahm Lincoln once said: "the best way to predict the future is to create it". Warner, however, also highlights the role that the Alliance has to play in strategy-making as technology is no substitute for strategy. As the age of intelligent machines is opening up, NATO with its member states should hence make sure to be at the forefront of both technology and strategy to rip the benefits of this transformation.

17 For a discussion on technological superiority, see A. Gilli and M. Gilli, "Why China has not caught up yet: military-technological superiority and the limits of imitation, reverse engineering, and cyber espionage", *International Security*, Vol.43, No.3, 2019, p.141-189.

1

An economic perspective on Artificial Intelligence

Ajay Agrawal, Joshua Gans, Avi Goldfarb[1]

Artificial intelligence (AI) technology has improved rapidly over the last decade, generating both excitement and concern, especially with regard to machines using artificial intelligence of a general nature. While such machines may one day be achievable, most people struggle to understand the essence and implications of artificial intelligence.[2] In our book, *Prediction Machines: The Simple Economics of Artificial Intelligence,* we argue that artificial intelligence, and in particular the field of computer science on which recent progress is based – machine learning – is best seen as a prediction technology.[3] In this way, we can better understand both its importance and its implications. Technological progress often makes certain tasks cheaper. It is useful to understand artificial intelligence as making prediction tasks more affordable. Such affordable technology, in turn, becomes more pervasive as new opportunities to use the technology arise. When technology becomes cheaper, demand also increases for its complements. In the case of prediction technology, a key complement is judgment, or the ability to understand payoffs. Future competition in AI will thus be about accessing, developing and possessing key complements like judgement. Last, but not least, technological change triggers organizational changes. This happened during the industrial Revolution with the steam engine and electricity: the same is likely to happen with artificial intelligence and machine learning. But organizational change is neither easy nor immediate – as the following chapters will further illustrate.

---

1 We thank Leah Morris for her research assistance.

2 D. Kahneman, "Comment on Artificial Intelligence and behavioral economics", in A. Agrawal, J. Gans, and A. Goldfarb (ed.), *The economics of Artificial Intelligence: an agenda*, 2019, pp.608-610, https://www.youtube.com/watch?v=gbj_NsgNe7A (accessed 19 April 2019).

3 A. Agrawal, J. Gans, and A. Goldfarb, *Prediction machines: the simple economics of Artificial Intelligence,* MA, Cambridge, Harvard Business School Press, 2018.

Technological change and cost

Artificial intelligence is generating enormous expectations as well as concerns. Some, like Elon Musk, believe that intelligent machines will be much smarter than human beings – and this may have significant albeit not necessarily positive implications for our life, including our freedom. Different scholars like Nicholas Bostrom and Ruy Kurzwel (currently at Google), have long argued that this path is somehow inevitable.[4] Several consulting companies like Accenture or PricewaterhouseCoopers (PwC) are much more optimistic, predicting that artificial intelligence will add several USD billion to the global economy in the near future.

The debate about artificial intelligence is, however, somehow nebulous. The primary technology underlying the current excitement in AI is machine learning. Machine learning can be understood as prediction technology. Prediction is the process of filling in missing information. It uses available information (or data) to generate unavailable or unknown information. This new information can inform predictions not only about the future but also about the present. For example, when a credit card company refuses a transaction because it suspects a card is stolen, it is predicting the present. Recent advances in artificial intelligence have led to more accurate, faster, and cheaper prediction. Computing, understood as a cost reduction in arithmetic, provides an example.[5] Computers are used for many tasks, but in reality they are just a tool for arithmetic: they handle numerous calculations at speeds beyond the capabilities of humans. Once arithmetic became cheap enough, we started to use it much more, also in fields or for purposes not initially envisaged.

Cheap arithmetic meant that these tasks could be performed at greater speed and accuracy than when carried out by human beings. Over time, the cost of arithmetic fell further. Another example is provided by the field of photography: until the 1990s, the problem was largely a chemical one. Kodak, in many ways, was a chemical company. As the price of arithmetic fell, however, it became feasible to treat photography as an arithmetic problem to be solved by computers. Similarly, music recording was a combination of chemistry and physics, solved by using magnets and various plastics. Nearly 200 years ago, Ada Lovelace discovered that many of these problems could be solved by arithmetic.[6] This happened when arithmetic became cheap enough, but only more recently.

4 See R. Kurzweil, *The singularity is near: when humans transcend biology*, New York, Penguin, 2006; N. Bolstom, *Superintelligence: paths, dangers, strategies*, First ed., Oxford, Oxford University Press, 2014.

5 W. D. Nordhaus, "Two centuries of productivity growth in computing", *Journal of Economic History,* Vol.67, No.1, 2007, pp.128-159.

6 W. Isaacson, *The innovators: how a group of hackers, geniuses, and geeks created the digital revolution*, Toronto, Simon & Shuster, 2014.

Thus, as arithmetic has become cheaper, we have used more arithmetic in applications that were not widely seen as arithmetical problems. This relates to one of the core concepts of economics: when prices fall, the demand for quantity rises. In other words, cheap arithmetic means more arithmetic because demand curves slope downward.

The early applications of machine learning (i.e. machine prediction) were well-established (human) prediction problems. As technology has improved and has become more accurate, and prediction has become cheaper, new applications for prediction have arisen. Machine learning is in fact already widely used in inventory and logistics. It helps predict when inventories might run short so that new shipments can be ordered. It is also used to predict delivery times in supply chains. Many companies are already using machine learning to predict which job applicants are likely to perform best.[7] The technology could also be used in recruitment; predicting who in the general population is likely to be a good fit for a particular organization, whether that organization is a university, a company, or the military. In some contexts, cheaper prediction has also enabled medical diagnosis and language translation to be solved using prediction technology. The fact that technology may offer these opportunities does not mean, however, that the transition will be smooth.

Substitutes and complements

In the simplest terms, we demand more of a product or service as its price falls. When coffee is inexpensive, we buy more coffee. Similarly, when prediction is inexpensive, we do more prediction. However, an increase in demand affects other products and behaviours. Take the following simple example: cheap coffee also means we buy less tea. This is because tea and coffee are substitutes. Therefore, as the price of coffee falls, the amount of tea purchased also falls. Tea and coffee are generic examples that, however, help us understand why people have started flying much more often when Ryanair offered its very low fares throughout Europe or why with the emergence of Uber and other car-sharing services, people have started buying, or using, cars less. This logic brings us to our discussion of the economics of artificial intelligence: the main substitute for machine prediction is human prediction. For this reason, prediction tasks currently done by humans will increasingly be done by machines, whether those tasks are medical diagnosis, recruitment, or inventory management.

7 B. Cowgill and C. Tucker, "Economics, fairness, and algorithmic bias", *Journal of Economic Perspectives*, Forthcoming, 2019.

Cheap coffee also means we buy more cream and sugar. Similarly, the growth of Ryanair's business called for the the rise of new transportation services to and from airports as well as the expansion of many of those very airports – whose existing facilities could not handle the increase in passengers and air traffic. Cream and sugar, like transport services and airport facilities, are complements. The largest opportunities with respect to machine prediction require investment in complements. In order to identify the complements to prediction, it is important to understand why prediction is valuable. A prediction alone has no value. Predictions are only valuable because they improve decision-making. Therefore, the complements come from the other facets of decision-making: data, action, and judgment.

In *Prediction Machines,* we discuss how these three complements are affected.[8] First, data is a key input in prediction. More data enables better predictions. Data becomes more valuable and the ability to collect data in the future increases in value. Second, predictions have little use if no action can be taken. The actions taken in any context can be more effective in the presence of better prediction. There are, however, a set of challenges related to these two issues, including about bias in data, privacy as well as ethical questions about decisions. They are further discussed in Chapter 4. What matters for this discussion is that their implications are also economic and operational, not just legal and ethical.

The final, and perhaps most important, complement is judgment. Judgment is the ability to understand payoffs. It is the reward for action taken in a given environment. A prediction alone is not useful in the absence of judgment. The movie *I, Robot* provides a useful example. In the movie, Will Smith plays the protagonist. There is a flashback scene that explains why his character hates robots. He and a 12-year-old girl are trapped in the water after a car accident that causes them to swerve off the road and into a river. Just as it seems that both of them are about to drown a robot appears and saves the adult, but not the child. After the accident, Will Smith's character figures out why he was saved and not the girl. In auditing the robot's decision, he discovers that the robot made predictions about the likelihood of survival. It predicted that he had a 45 percent chance of survival while the girl only had an 11 percent chance. Given those predictions, the robot saved him and not the child.

In his view, that was the wrong decision and a human would have known to save the girl. This assessment says something very particular about judgment. It implies that the girl's life is worth more than four times his life. The prediction alone is not enough to make

8 This framing builds on decades of research into decision theory, summarized for example in I. Gilboa, *Making better decisions: decision theory in practice,* Hoboken, Wiley-Blackwell, 2010.

a decision about who to save. Decisions also require judgment on the relative value of the different outcomes. Recent advances in artificial intelligence are advances in prediction, not judgment. It is still necessary for humans to tell the machines what to predict, and what actions to take given the predictions, but this requires a broader ethical discussion (see Chapter 4).

Prediction technology in organizations

The vast majority of prediction technologies will be applied as tools for particular tasks. Within any organization's workflow, there are a large number of tasks. Some of those tasks can be seen as prediction tasks. As prediction machines get better, machines will be doing those tasks instead of humans. Whether the machine is doing inventory management, medical diagnosis, or recruitment, in most cases the overall structure of the organization will not change. As companies use machine learning to help them screen resumés, new recruits still arrive at their jobs and work. It is just that the hiring process could be more efficient and more effective.

In this sense, automating tasks may not decrease the need for humans in the workflow. It may even increase the demand for some human skills. Both in his book, and in his chapter in this volume, Steven Fino explains how the use of guided missiles in F-4 fighter jets led to a need for a second human in the cockpit in order to operate the radar.[9] Similarly, improved prediction technology may sometimes enable humans to be more productive, without reducing the amount of work they do.

Sometimes, improved prediction can transform the way an organization works. In *Prediction Machines* we examine Amazon as an example. Currently, Amazon makes predictions about what products their customers want to buy. These predictions are impressive, considering the millions of products in Amazon's catalogue, they often make recommendations that result in a purchase. McKinsey estimates that 35 percent of Amazon purchases come from these product recommendations.[10] These predictions are tools that Amazon uses to help its customers complete tasks more efficiently. In this case, predictions help find what the customer wants. The tools fit within the current business model, which in many ways is similar to what Sears and other mail-order catalogue companies did in the

9 S. A. Fino, *Tiger check: automating the US Air Force Fighter Pilot in air-to-air combat, 1950-1980*, Baltimore, Johns Hopkins University Press, 2017.

10 I. MacKenzie, C. Meyer, and S. Noble, *How retailers can keep up with consumers,* Minneapolis, McKinsey & Company, October 2013, http://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers (accessed 19 April 2019).

late 19th century. Amazon has an online catalogue, customers look through it and choose what they want, and then Amazon sends the product from their warehouse to the customer. They are much more efficient at this than the old catalogue companies, but the business model is not fundamentally different.

If Amazon's predictions get much better, the business model may be transformed. Suppose Amazon could predict what products a customer will want with much greater accuracy. If their predictions are accurate enough, they will not wait for the customer to order from the catalogue. Instead, Amazon will ship the item directly to the customer before it is ordered. The predictions do not need to be perfect, they only need to be good enough for it to be more profitable to build an infrastructure for dealing with product returns. Put differently, better prediction could transform the way Amazon runs its logistics – but also the broader online shopping market, with important implications for competition and market access.

Military recruiting is another example. The current process is time-consuming, challenging, and expensive. A key challenge is the identification of potential recruits who will be interested in enlisting and will successfully progress through the various steps from the fitness exam, the training, the first term, and potentially reenlistment.[11] A recent 10-year contract for the United States Army's advertising campaign reportedly approached USD4 billion.[12] Better prediction could lead to reductions in the costs of recruitment. In the short term, it could enable better-targeted advertising and incrementally improve the quantity and quality of recruits. In the long-term, if the predictions improved sufficiently, it could lead to personalized recruiting messages that could bypass mass media entirely.

Medicine may also be transformed. Doctors do a variety of tasks including diagnosing conditions, prescribing and administering treatments, collecting patient information, and explaining procedures and results with patients.[13] One key skill that distinguishes doctors from other medical personnel is the ability to diagnose – among many others. In other words, in many contexts, a medical doctor is needed – rather than a nurse, technician, social worker, or paramedic – because the doctor is trained in diagnosis.

11 For example, see R. J. Buddin, "Success of first-term soldiers: the effects of recruiting practices and recruit characteristics", RAND Corporation, 2005, Monograph.

12 P. Coffee, *Omnicom wins 10 year, $4 billion army contract after elimination of incumbent McCann: DDB Chicago led the pitch*, New York, Adweek, LLC, 20 November 2018, http://www.adweek.com/agencies/omnicom-wins-4-billion-army-review-after-elimination-of-incumbent-mccann/ (accessed 19 April 2019).

13 O*NET OnLine, *Summary Report for: Family and General Practitioners,* O*NET, 2019, https://www.onetonline.org/link/summary/29-1062.00 (accessed 19 April 2019).

Diagnosis can be seen as a prediction task. Your doctor takes in information about your symptoms and fills in the missing information about the cause of those symptoms. Diagnosis is the process of filling in missing information. It is prediction. If machine prediction continues to improve, then many tasks that currently require a medical doctor will instead require different skills. Nurses, technicians, social workers, and paramedics may be able to conduct tasks previously reserved for personnel with medical degrees.

The medical example suggests that better prediction can also change power structures. In many healthcare organizations, the doctor's ability to diagnose represents a power differential between doctors and other medical professionals.[14] While doctors do many things in their workflow, much of the selection into medical school and the training of doctors focuses on diagnosis. It is an important source of the doctor's wage premium. If the doctor's expertise in diagnosis has less value for patient outcomes, this could change the power relations between doctors, nurses, and others in the delivery of healthcare. This conjecture provides a useful example of how technological change changes the relative value of skills, in turn affecting the relative prices of production factors and thus the bargaining power of different categories of actors.

Prediction and power

Machines do not make decisions. People do. Artificial intelligence enables the decisions to be moved in time and space. By moving the decisions in time and space, the decisions become probabilistic. Rather than one at a time, the person making the decision specifies some broader framework that determines the distribution of future outcomes, rather than a specific outcome.

In a military context, many decisions are delegated to individual soldiers. This enables militaries to operate at scale efficiently around challenges such as geographic separation, constant team reformation, and unstable operational contexts. Due to these uncertainties, individual decision-making is a strategic choice made by militaries to leverage a soldier's ability to predict, given that they likely have the most information at any given time. However, better prediction could reduce uncertainty, shifting the decision-making power away from the deployed soldier toward higher-ranking individuals. Whether this would be a solution to the problem of tactical Generals, emerged because of the opportunities offered

14 S. Porter, "A participant observation study of power relations between nurses and doctors in a general hospital", *Journal of Advanced Nursing,* Vol.16, 1991, pp.728-735.

by instant communication, or just exacerbate it is, however, another discussion.[15]

For instance, the lowest-ranking non-commissioned officer (NCO) in the US Marine Corps is responsible for developing, conducting, and evaluating training for their subordinates.[16] After a minimum of two years of service, the corporal will be responsible for small teams of three to nine individuals,[17] including evaluations and other feedback such as *Performance Evaluation Checklists* that help to assess the training of soldiers and the unit as a whole.[18] Ultimately, the commander will use this input to make a judgment about the unit and make deployment decisions.[19]

At this time, NCOs provide, among other things, the best prediction on how a soldier will perform when deployed, therefore they have the power to decide which soldiers meet the training standards.[20] Better prediction could mean that commanders will no longer depend on NCOs for the necessary input to decide which soldiers meet training standards. In this case, better prediction moves the decision about training standards up the chain of command. The caveat, however, is the word judgement in the previous paragraph: commanders will have to perform as well as NCOs in judgement without direct personal relations and observation.

This process could happen in a variety of contexts. Better prediction could reduce the need to delegate decisions, and increase the power and control of the higher-ranking members of any organization, military or otherwise.

The other chapters in this volume show some of the broad challenges that the current wave of technological transformation raises. In this chapter, we have summarised the longer argument presented in our book, namely that by interpreting recent advances in artificial intelligence as improvements in prediction technology, we can understand and clarify many

15 Peter Singer just raised attention to the risk that, with real-time communication, Generals would try to take over the responsibilities of theatre commanders. In his view, this risked undermining military effectiveness, consistently with the research from Stephen Biddle. See S. D. Biddle, *Military power: victory and defeat in modern warfare*, Princeton, Princeton University Press, 2004, and. P. W. Singer, *Robots and the rise of "tactical Generals",* Brookings, 9 March 2009, https://www.brookings.edu/articles/robots-and-the-rise-of-tactical-generals/ (accessed 10 October 2019).

16 R. Barcott, "The strategic corporal, leadership lessons from the military", *Harvard Business Review*, 21 October 2010, http://hbr.org/2010/10/the-strategic-corporal.html (accessed 14 May 2019).

17 *Ibid.*

18 Commandant of the Marine Corps, *Marine Corps Combat Readiness Evaluation (MCCRE),* Washington, DC, Department of the Navy, 25 February 2019, http://www.marines.mil/Portals/59/Publications/MCO%20 3501.1E%20%20v2.pdf?ver=2019-04-01-135547-403 (accessed 14 May 2019).

19 *Ibid.,* pp.4-9.

20 G.B. Stupp, *The staff noncommissioned officer's handbook,* Fort Leavenworth, 2015, p.30.

– although not necessarily all – of its implications and consequences. More specifically, as costs of prediction technology fall and accuracy increases, technology becomes more pervasive and thus used also for purposes not initially conceived. Improvements in prediction enhance the productivity of existing processes. However, they also require some complements such as data, action and judgement. Accessing, developing or possessing such complements is challenging. Some of these issues are further explored in Chapter 4 on AI and ethics as data can be biased, and thus lead to both unethical and uneconomic predictions. Sometimes, however, by reducing uncertainty, the reduction in the cost of prediction can transform the strategy and structure of an organization. But as Chapters 2 and 3 illustrate, such changes are far from easy due to psychological, cognitive or political reasons. In conclusion, these considerations suggest that in the future, competition will be shaped by access to complements as well as by strategy-making capabilities and organisational flexibility. As prediction technology becomes more easily available, other factors – mostly tightly related to human beings – will determine success and failure as well as victory and defeat.

2

Human psychology and intelligent machines

Aaron Celaya and Nick Yeung[1]

The cheaper artificial intelligence (AI) and its related technologies become, the more they will become pervasive: they will be employed outside of the domains for which they were initially conceived. As the previous chapters have illustrated, automation will take over more mundane and repetitive tasks, thus making the role of human beings, paradoxically, more important in the decision process. As a result, the benefits of artificial intelligence will depend, critically, on the interaction between humans and machines, first, and on their acceptance. These two topics have little to do with the technology and much to do with the psychology of human beings. Machines can be, in theory, perfectly designed, but they will be of little utility if users have pratical biases in using or trusting them. Similarly, it will be difficult to benefit from intelligent machines if they are not accepted – not an unlikely outcome if we look at the historical record.[2] This chapter provides an introduction to the literature on the psychology of the interaction between human beings and machines, illustrating what aspects practitioners will have to consider – at the tactical, operational and strategic levels.

Artificial intelligence and human psychology

Research, development, and deployment of AI systems are increasing across the whole of society, in contexts as diverse as banking and finance, education, cyber defense, marketing, gaming, medical diagnosis, and security. AI systems are no less on the rise in warfare environments, where we are seeing development in military contexts ranging

1 The views expressed in this NDC Research Paper are the responsibility of the authors and do not necessarily reflect the opinions of the NATO Defense College, the North Atlantic Treaty Organization, or any other institution represented by the contributors.

2 C. B. Frey, *The technology trap: capital, labor, and power in the age of automation*, Princeton, Princeton University Press, 2019.

from targeting to intelligence exploitation, battlefield medical response, bomb diffusion, subterranean mapping, and automated Observe-Orient-Decide-Act (OODA) Loop inputs. The increasing influence of AI is widely expected to have a transformational, rather than incremental, impact on military operations. Former NATO Deputy Secretary General Rose Gottemoeller underscored the criticality and urgency of AI when noting that "we cannot fight tomorrow's threats with today's tools […] in the age of artificial intelligence".[3]

However, notwithstanding the many striking successes of AI in recent years – from mastering the game of Go to AI-aided disaster relief response – many of these systems or concepts are still in their infancy. In particular, AI capabilities that have matured to date, even those designed to learn from their experiences rather than execute pre-specified computations, are nevertheless specialized to perform certain, well-defined functions and subsequently remain limited in scope and application.[4] The most immediate goal is thus to develop AI that can mimic the core cognitive capabilities of children, 0-18 months old. The fact that this goal is genuinely ambitious highlights the challenge of developing AI systems that are capable of usefully contributing in contexts requiring judgment, flexibility, and the ability to make informed decisions on the basis of sparse information – properties that characterize most, if not all, political and military contexts.

Two related implications follow from this analysis: first, for the foreseeable future, AI systems will often be deployed in conjunction with human operators and decision makers, rather than acting autonomously. Second, the effectiveness of AI systems will depend critically on their usability and acceptability to humans. For this reason, "realising this potential will depend on understanding the relative strengths of humans and machines, and how they best function in combination to outperform an opponent. Developing the right blend of human-machine teams – the effective integration of humans and machines into our war fighting systems – is the key".[5]

As such, an understanding of human psychology will play a critical role in maximizing the potential of intelligent machines. At a basic level, the foundation of effective human-machine teaming will lie in identifying the respective strengths and weaknesses of human and machine intelligence. The 2018 US Department of Defense's Artificial Intelligence

3 Remarks by NATO Deputy Secretary General Rose Gottemoeller at the Xiangshan Forum in Beijing, China, 25 October 2018, NATO Newsroom, Speeches & Transcripts Section, https://www.nato.int/cps/en/natohq/opinions_160121.htm (acessed 8 November 2019).

4 D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search", *Nature*, No.529, January 28, 2016, pp.484-489.

5 *Joint Concept Note 1/18: human-machine teaming*, UK Ministry of Defense, May 18, 2018, https://www.gov.uk/government/publications/human-machine-teaming-jcn-118 (accessed 8 November 2019).

strategy, for instance, directs the use of AI "in a human-centred manner" that recognizes the complementary strengths and weaknesses of human and machine intelligence, with the goal of simplifying workflows and improving the speed and accuracy of repetitive tasks in order "to shift human attention to higher-level reasoning and judgment, which remain areas in which the human role is critical".[6]

However, as AI systems continue to develop in sophistication and reach, they will likely – and perhaps inevitably – move beyond the status of tools to be deployed by people, and instead take on roles as peers (or even the superiors) of human operators: setting goals, developing strategies, drawing inferences, and making meaningful decisions in domains that are currently the exclusive province of human intelligence. As such, with AI systems acting as partners to humans – true human-machine teaming – we foresee a central role for understanding the psychology of trust, teamwork, and communication in effective AI system design.

Human-machine trust

Teamwork relies on trust: the willingness of individuals to work interdependently, such that they accept roles in which they are vulnerable to the actions of other team members with the expectation that those others will act accordingly.[7] Existing research on human-machine teaming highlights the complexity of optimizing human users' trust in artificial systems. This challenge is particularly evident when errors in the system have been observed by the human user, with evidence of these users veering from irrational aversion to irrational overtrust in artificial systems.

Consistent with a longstanding observation that, though statistical prediction outperforms clinical or human-only prediction in many situations, people still prefer to get advice or diagnoses from other people rather than expert systems,[8] recent experimental research found that human decision makers consistently downweight information provided

6 *Summary of the 2018 Department of Defense Artificial Intelligence Strategy*, US Department of Defense, 2018, https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF (accessed 8 November 2019).

7 R.C. Mayer *et al.*, "An integrative model of organizational trust", *Academy of Management Review*, Vol.20, No.3, 1995, pp.709-734.

8 P. Meehl, *Clinical versus statistical prediction: a theoretical analysis and a review of the evidence*, Minneapolis, University of Minnesota Press, 1954; W. Grove *et al.*, "Clinical versus mechanical prediction: a meta-analysis", *Psychological Assessment*, Vol.12, No.1, 2000, pp.19-30.

by intelligent machines such as expert systems and algorithm-based prediction systems.[9] For example, in one study, participants were asked to forecast the success of applicants to an MBA program (based on the information available in their file), and were then given the option of using forecasts from another person (a previous participant in the experiment) or a computer algorithm (a statistical prediction model). Absent any experience with forecasts of either the model or the past participant, most people preferred to use the algorithm's forecasts. However, when given experience with the statistical prediction model, or with both the model and the other person's forecasts, participants were consistently less likely to choose the model, even when the algorithm clearly outperformed both the participant themselves and human advisor. Apparently, such "algorithm aversion" arises because people rapidly lose trust in algorithms after seeing them err, but seem much more forgiving of human error.

Other works, however, have documented another problem, namely dramatic cases of overtrust in artificial systems.[10] Overtrust occurs when too much risk is accepted by the human user with the belief that the autonomous or AI counterpart will make up the difference and mitigate the accepted risk. In some recent studies, participants were willing to trust an emergency guide robot, even when they had observed it malfunction previously. In contrast to the works discussed before, human users were apparently very quick to forgive a robot for previous mistakes and to trust them even to the point of failing at their overall objective. One explanation is that users did not possess an accurate assessment of the robot's capabilities and accuracy. Another explanation holds that people just followed the robot's advice/directions simply because the robot stated that it was designed for the task, thus suggesting that designation mattered more than performance.

These studies identify divergent patterns of aversion versus overtrust, but converge in showing that human agents can learn incorrect models of machine trust and therefore use highly sub-optimal machine reliance strategies. The implication is straightforward: developing AI systems that can be trusted appropriately by human operators is of utmost importance. Our working hypothesis is that this challenge can be addressed by applying established principles of trust in normal human-to-human interactions. The concepts in human relationships form the foundation for those applied in human-machine teams.[11] For

9 B. Dietvorst *et al.*, "Algorithm aversion: people erroneously avoid algorithms after seeing them err", *Journal of Experimental Psychology: General*, Vol.144, No.1, 2014, pp.114-126.

10 P. Robinette *et al.*, "Overtrust of robots in emergency evacuation scenarios", *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Christchurch, 2016, pp.101-108.

11 J. Jian *et al.*, "Foundations for an empirically determined scale of trust in automated systems", *International Journal of Cognitive Ergonomics*, Vol.4, No.1, 2000, pp.53-71.

example, human decision makers exhibit egocentric bias, discounting advice from other people relative to their own opinions. Further, human decision makers are quick to form opinions of their advisors which are updated asymmetrically: trust is built slowly, but lost rapidly when advice turns out to be incorrect.[12] Human decision makers are also likely to be influenced more by advice they request and to discount advice less when the task is complex. Alternatively, human decision makers discount advice more as the difference between their initial judgment and the advisor's opinion increases.[13]

In our research, we apply such findings and theories of human trust as our framework for exploring trust in human-machine teams. A cornerstone of our approach characterizes the relationship between *trustor* and *trustee* as depending jointly on properties of each individual and the interaction between them (see Figure 1). The trustee is characterized by three traits:

- ability – the skills and competence of the trustee in their assigned tasks;
- benevolence – the extent to which the trustee is believed to want to do good to the trustor; and
- integrity – whether the trustee adheres to a set of principles that the trustor finds acceptable.

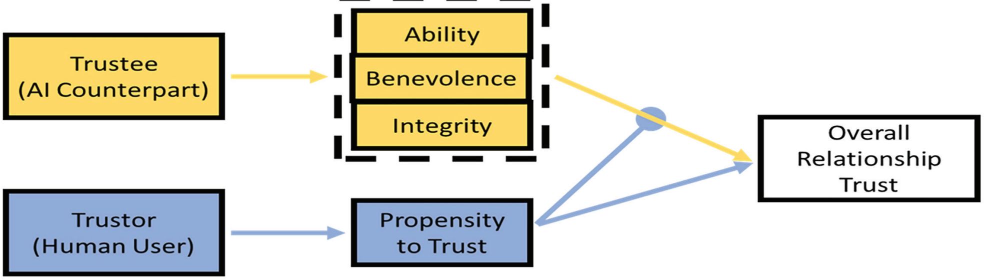


Figure 1: Trust model adapted from Mayer *et al.* (1995)

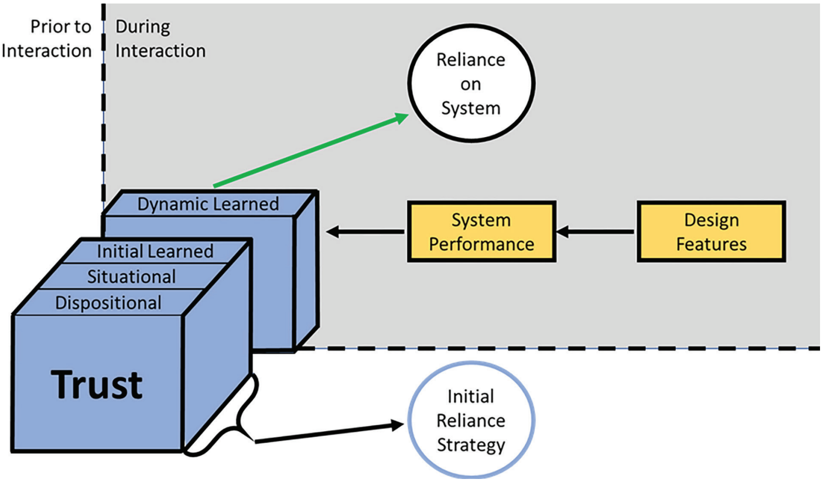
12 I. Yaniv and E. Kleinberger, "Advice taking in decision-making: egocentric discounting and reputation formation", *Organizational Behavior and Human Decision Processes*, Vol.83, No.2, 2000, pp.260-281.

13 S. Bonaccio and R. Dalal, "Advice taking and decision-making: an integrative literature review, and implications for the organizational sciences", *Organizational Behavior and Human Decision Processes*, Vol.101, No.2, 2006, pp.127-151.

The trustor, meanwhile, is characterized by the propensity to trust: individuals may vary in their baseline levels of trust in others, and also more subtly in the relative weight they place on ability, benevolence and integrity as critical to winning and retaining their trust.

Hoff and Bashir propose that trust is simpler in the context of human interactions with machines, being determined only by the ability of the machine and the human's propensity to trust (i.e., with benevolence and integrity of the machine being less important – a point we return to below).[14] Within this simplification, their model usefully elaborates how individuals vary in their propensity to trust: this propensity depends on the individual's general levels of trust in others (*dispositional trust*) that varies according to variables such as culture, age, gender or personality, as well as more context-specific factors (*situational trust*) relating to the current state of the person (e.g., their mood, their current workload), and to the situation in which they find themselves (e.g., the particular system they are working with, the complexity of the task being solved). Together, these dispositional and situational factors form an initial reliance strategy (see Figure 2). This strategy is then modified through interaction according to the performance and behavior of the machine (synonymous with ability in Roger C. Mayer *et al.*'s model), specifically how well it performs its given task. This dynamic learned trust continues to evolve over the course of system interaction and determines the trustor's reliance on the system.

Figure 2: Trust model adapted from Hoff & Bashir (2015)



14 K. Hoff and M. Bashir, "Trust in automation: integrating empirical evidence on factors that influence trust", *Human Factors,* Vol.57, No.3, 2015, pp.407-434.

Current research efforts

We are conducting experimental work within these theoretical frameworks of trust, exploring human-machine teaming in tasks with requirements similar to those of operationally-relevant perceptual judgments (e.g., "Is there a surface-to-air missile site in this piece of satellite imagery?"; "Are there subtle changes over time in this region of interest?"). Our aim is to more fully understand trust in AI systems, as they depend on individual user differences (i.e., propensity to trust) and the properties of the AI systems (i.e., ability). In the experiments, human participants make a series of visual judgments, with the option of receiving and using advice from another person (a previous experimental participant) or a simple AI system (computer algorithm). We measure trust in terms of our participants' choice of advisors and the relative influence that advisors have on participants' ultimate decisions.

Figure 3: Correlation of advisor influence and choice

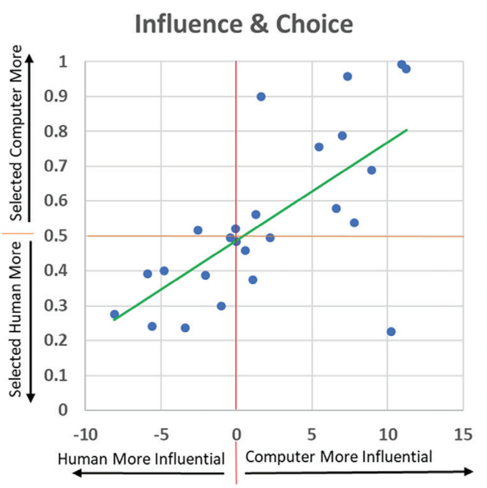


Figure 3 presents illustrative findings from one of our experiments, in which we controlled for system performance, design features, and initial learned information – all participants saw the same advice quality from the two (human and computer) advisors – in an effort to isolate and study individual differences in propensity to trust. In the figure, each point corresponds to one of our experimental participants, with the x-axis value indicating whether each participant was more influenced by human vs. computer advice, and the y-axis value indicating whether they were more likely to choose human vs. computer advice. The results reveal neither systematic aversion nor overtrust in AI, in contrast to the studies discussed before: the data points are evenly distributed either side of the red neutral midpoint, and above/below the orange neutral midpoint of each axis. However, large idiosyncratic individual differences are clearly apparent. Even though human and computer advice was objectively equally accurate, some participants were more influenced by the human advisor and chose him/her more (data points clustered to the lower left of the figure), whereas other people were more influenced by and chose the computer advisor more often (data points clustered in the upper right). Therefore, our participants are rational insofar as they choose the advisor that most influences them, but are characterized by substantial (irrational) variation in their preference for one advisor over the other. Leveraging these findings, we are now conducting experiments to understand what dispositional and situational factors drive these preferences with a view to developing profiles of individuals' trust in AI.

Other experiments have allowed us to characterize the dependence of human-machine trust on observed ability of the system. For example, in line with evidence that human-human trust is rapidly lost and only painstakingly rebuilt after bad advice, we observe that early experience with advisors is critical when determining appropriate reliance strategies (Figure 4).[15] In this experiment, the average accuracy of the human and computer advisors was perfectly matched across the whole experiment. However, from trial to trial there is natural variability in advisor ability as perceived by the participant, reflecting the uncertainty inherent in difficult judgments. Variability in the first interactions with advisors turned out to have long-lasting effect on participants' trust. If a participant's initial experience was of the human advisor being more accurate than the computer, they tended to choose that advisor for the remaining interactions (data points clustered to the lower left of the figure), and the same was true if initial interactions favored the computer advisor (data points towards the upper right). This over-reliance on one advisor resulted in better perceived performance by that advisor, as predicted by Kevin A. Hoff and Masooda Bashir.[16] Translated into practical implications, these findings indicate that training human operators on new systems might benefit from demonstrating system performance via simulation and classroom instruction. Without this component, the human operator could form an initial reliance strategy based on unrepresentative early experiences that are different from what will be developed later, after real-world system interaction.

15 I. Yaniv and E. Kleinberger, *Advice taking in decision making*, pp.260-281.

16 K. Hoff and M. Bashir, *Trust in automation*, pp.407-434.

Figure 4: Correlation of early experience and choice

Outlook

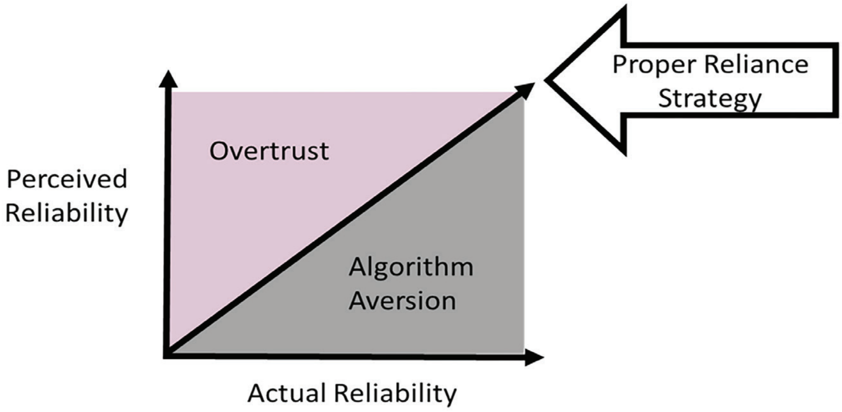
To the extent that AI systems remain embedded in human-machine teams rather than acting fully autonomously, understanding human psychology will be critical to maximizing the potential of these emerging technologies. Here we have focused on one example: applying principles of human trust to ensure appropriate reliance on machines. Within Keith Gempler's[17] depiction of the relationship between perceived and actual system reliability (Figure 5), we have seen how human operators deviate from proper reliance strategies as a function of their disposition, pre-conceptions and experiences, such that they exhibit aversion (underestimation of machine reliability) or overtrust (overestimation of reliability) in machine partners. These findings indicate the importance of system designs that mitigate against these human biases, which can again be informed by known psychological principles of trust maintenance and repair. For example, in human teams, communication of confidence and uncertainty is crucial to both effective group decision making and trust calibration, suggesting that effective human-machine teaming may depend on the development of AI systems that know (and signal) when they might be wrong.[18]

---

17 K. Gempler, *Display of predictor reliability on a cockpit display of traffic information*, Urbana, University of Illinois at Urbana-Champaign, 1999.

18 B. Bahrami *et al.*, "Optimally interacting minds", *Science*, Vol.329, No.5995, 2010, pp.1081-1085; N. Pescetelli and N. Yeung, *The role of decision confidence in advice-taking and trust formation*, 2019.

Figure 5: Reliability diagram adapted from Gempler (1999)



Looking further into the future, we foresee the importance of looking beyond simplified characterization of human-machine trust as dependent on human propensity to trust and machine ability alone,[19] and instead embrace a broader conception of trust that includes benevolence and integrity as key characteristics.[20] Thus, as AI systems grow in flexibility and autonomy, by necessity they will come to set goals and develop strategies that are independent of the original intentions of the system designers (as we are already seeing in David Silver *et al.*'s algorithm for the game Go, which taught itself super-human strategies that comprehensively outmaneuvered the very best human players of the game). Moreover, the goals and strategies developed by AI systems may be imperfectly aligned with those of their human partners. As AI systems become intentional agents in this manner, the benevolence of these systems towards human team members, and their integrity in terms of operating by principles that can be understood and endorsed by humans in the loop (as subordinates, peers and controllers of these systems) will be critical factors in assuring appropriate levels of trust and control in human-machine teaming.

19 K. Hoff and M. Bashir, "Trust in automation: integrating empirical evidence on factors that influence trust", *Human Factors,* Vol.57, No.3, 2014, pp.407-434.

20 R.C. Mayer *et al.*, "An integrative model of organizational trust", *Academy of Management Review*, Vol.20, No.3, 1995, pp.709-734.

3

Automation in air warfare: lessons for Artificial Intelligence today

Steven Fino[1]

Popular discussions regarding the application of Artificial Intelligence (AI) to military aviation frequently adhere to the standard "killer robot" trope. While images of a Terminator-like, terrorist-hunting drone may be useful for capturing our attention or igniting a philosophical debate about AI, more often they are an unfortunate distraction. By focusing exclusively on the AI *in* the drone, we tend to overlook the automation *of* the drone. There is value in looking beyond the specific AI-tools and instead focusing attention on the human decisions that determined what functions of the drone needed to be automated and why. This chapter therefore attempts to reframe and elevate today's conversations about AI by invoking the rich, socio-technical history of automation and providing an example of this alternative perspective. Specifically, reviewing the automation of air-to-air combat post-World War II reveals how different groups of operators, engineers, and their respective organizations developed and then responded to their newly-automated tools.

The previous two chapters made three main abstract or theoretical claims pertinent to this discussion. First, that technological progress triggers an increasing demand for complements. Next, that technological change often demands organizational change and a reallocation of tasks between human and machines. And, last but not least, that new technologies and organizational changes raise different psychological, sociological, and political reactions as the new tools and new roles emerge. The history of automation in air combat documented in this chapter illustrates these very dynamics. Ever increasing automation did not eliminate skilled pilots from the cockpit, but it did eliminate some of their historic tasks. Consequently, tasks had to be renegotiated and reallocated between the human pilots and their aircraft machines. Finally, the introduction of automation

1 The views expressed in this document are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the US Government.

generated different forms of resistance, opposition, and later acceptance withn the various costituences. Knowing and understanding this history is important because it can aid us today as we struggle to develop and adapt to the next generation of automation powered by the latest AI algorithms.

The three myths of automation

Developing novel technologies in the hope of automating once-burdensome human tasks is not a new phenomenon. Nonetheless, we do not have a stellar record of correctly anticipating the challenges that might accompany the automation. In his historical survey of various autonomous technologies spanning sub-surface, terrestrial, aviation, and space applications, David Mindell, a professor at the Massachusetts Institute of Technology (MIT), identified three pernicious myths of autonomy that have frequently distorted our understanding of automation's effects on society. The first, the *myth of linear progress*, reflects the popular belief that automation follows a natural, deterministic trajectory, whereby our machines steadily transition from requiring direct human involvement to the intermediary of remote human presence before finally becoming fully autonomous robots. Meanwhile, the *myth of replacement* underscores the presumption that the purpose of automation is simply to mechanize existing human tasks. Finally, the *myth of full autonomy* captures the utopian idea that robots can operate entirely on their own even after they are released into the messy, complex human world outside a laboratory.[2] A cursory review of the popular literature touting future AI-fueled automation, be it autonomous passenger vehicles operating on city streets or armed, expendable drones swarming a battlefield, illustrates how these three myths continue to shape our expectations.[3]

To move beyond these myths of autonomy, we must cultivate a rich understanding of the dynamic relationship that exists between humans, their tasks, and their equipment. These three components never operate in isolation, and they must therefore be studied as a whole as part of a *cognitive enthnography* – a "system of person-in-interaction-with technology".[4] Applying this analytic approach to the development of US Air Force fighter aircraft and their associated automated fire control technologies yields valuable insights for today.

2 D.A. Mindell, *Our robots, ourselves: robotics and the myths of autonomy,* New York, Viking, 2015, pp.8-10.

3 J.M. Lutin, "Not if, but when: autonomous driving and the future of transit", *Journal of Public Transportation*, Vol.21, No.1, 2018, pp.92-103; P. Scharre, "How swarming will change warfare", *Bulletin of the Atomic Scientists*, Vol.74, No.6, 2018, pp.385-389.

4 E. Hutchins, *Cognition in the wild,* Cambridge, MIT Press, 1995, p.155. See also T.B. Sheridan, *Telerobotics, automation, and human supervisory control,* Cambridge, MIT Press, 1992.

Patterns of automation in US Air Force fighter aviation

The United States Air Force (USAF) staked its future on advanced technology and automation even before its founding as a separate service in 1947. Exiting World War II, the then-Army Air Forces chartered its landmark, multivolume *Toward New Horizons* study to explore the relationship "between science and aerial warfare".[5] This early fascination with technology played a significant role in the fledgling service's later development, yielding what Carl Builder termed an "Icarus Syndrome".[6] Still, as one recent scholar noted, unlike the other US military services, the USAF is necessarily dependent on advanced technology to justify its very existence – humans do not fly without technology.[7]

From the F-86 Sabre to the F-4 Phantom to the F-15 Eagle and beyond, historic fighter pilot tasks have succumbed to ever-expanding aircraft automation. But the automation did not come conflict-free. Battles over who decided what tasks to automate and how to automate them were common, especially within the hyper-masculine community of jet fighter pilots. Over the last fifty years, those who flew and those who built the newest fighter aircraft have wrestled with a fundamental question: what is the goal of the automation? Is it to *eliminate* human skill in the cockpit or to *free* it to accomplish other tasks?

The F-86E Sabre

The F-86E Sabre was one of the first USAF fighter aircraft to feature a small radar to aid the pilot with aerial gunnery. Determining the range to an enemy target had always been a challenging prerequisite to a successful gunshot. Prior fighter pilots relied on a manually actuated, mechanical range estimator in their gunsight to guide their aim.[8] The process was cumbersome and pilots admitted it was especially challenging "trying to do everything at once" during a frenetic dogfight.[9] One engineer labeled the problem of range estimation

5 T. von Karman, "Science, the key to air supremacy: a report to General of the Army H.H. Arnold", in *Towards New Horizons*, Vol.1, Headquarters, Army Air Forces, 1946.

6 C.H. Builder, *The Icarus syndrome: the role of air power theory in the evolution and fate of the US Air Force,* New Brunswick, Transaction Publishers, 1994.

7 J.W. Donnithorne, *Four guardians: a principled agent view of American civil-military relations,* Johns Hopkins University Press, Baltimore, 2018, chapter 5. See also T.P. Schultz, *The problem with pilots,* Johns Hopkins University Press, Baltimore, 2018.

8 *Harmonization and firing techniques for fighter aircraft*, Armament Laboratory, Air Research and Development Command, March 1952, National Museum of the USAF (NMUSAF), pp.31-37.

9 D.K. Evans, *Sabre jets over Korea: a firsthand account,* Blue Ridge Summit, Tab Books, 1984, p.85.

"the bugaboo of aerial gunnery".[10]

The USAF's new A-1CM gunsight, and its accompanying small APG-30 radar, promised to solve the problem. Forecast in *Toward New Horizons* and designed by a team at MIT, the new radar-equipped gunsight was hailed as a major innovation.[11] The *New York Times* boasted the new "secret radar sight [...] does everything but fly the plane".[12] USAF publications similarly extolled the virtues of a "gunsight almost completely automatic in its operation", proclaiming the A-1CM "superior to any other fixed Gun Fire Control system".[13] No longer would a pilot need to spend countless hours practicing his dogfighting skills and learning how to close with his adversary. Instead, now aided by the automated A-1CM, fighter pilots would be able to score hits "with extreme accuracy at hitherto impossible ranges".[14]

The automated gunsight heralded a new division of cockpit tasks, at least according to the engineers. The human obviously would still pilot the aircraft, but when it came to shooting down a foe, the gunsight would be in charge; the pilot would simply maneuver his aircraft to place the gunsight's automatically calculated aiming pipper over the target and then squeeze the trigger.[15] Reflecting this arrangement, the pilot was provided with only minimal controls for the gunsight and its accompanying equipment. In the case of the radar, that included only a single power switch, a single light bulb that illuminated whenever the radar achieved a radar lock, and a single button on the control stick to temporarily break the radar lock.[16] Engineers scolded stubborn fighter pilots who didn't adhere to the defined relationship, informing those pilots that "snap decision corrections or judging will

10 F.H. Greene, Jr., to Commanding General, Continental Air Command, 31 January 1950, Charles Stark Draper Lab-Historical Collection (CSDL-HC).

11 L.A. DuBridge *et al.*, "Radar and communications: a report for the AAF Scientific Advisory Group", *Toward New Horizons*, Vol.11, p.22.

12 "New radar sight guides jets' guns", *New York Times*, 3 April 1950.

13 *A-1 series (gun-bomb-rocket) sight,* Continental Air Command Manual 50-11, Vol.2, No.2, Mitchell AFB, Headquarters, Continental Air Command, 1952, Air Force Historical Research Agency (AFHRA); *Operational suitability test of the A-1CM gun-bomb-rocket sight (F-84E Phase with AN/APG-30 Radar Ranging),* Project APG/ADB/8-A, Eglin AFB, Air Proving Ground Command, 18 December 1950, Defense Technical Information Center (DTIC), p.2.

14 *Harmonization and firing techniques for fighter aircraft*, p.1; *Operational suitability test of the A-1CM gun-bomb-rocket sight*, p.10. Because fighter aviation in the USAF was closed to women until the early 1990s, male-specific pronouns are used in this text.

15 I.A. Getting, "Conference on fighter fire control", Ad Hoc Group on Airborne Fire Control Systems for Fighter Aircraft, Cambridge, MIT Instrumentation Laboratory, 16 January 1953, DTIC, p.I-5.

16 *Harmonization and firing techniques for fighter aircraft*, pp.42-44; *T.O.1F-86A-1*, F-86A Aircraft, Flight Handbook, 30 August 1957, NMUSAF, sect.4, pp.20-23.

not improve results obtainable with the A-1C sight".[17]

The fighter pilots dueling in the skies over the Yalu River during the Korean War had a different perspective. For them, the new electronic gadgetry had not changed dogfighting nor reduced their human role within it. "It's still a matter of seeing first and trying to get on the other's tail or shaking him off yours", one explained.[18] Another preached, "[a]pproximately 90 percent of your air-to-air combat is positioning. [...] If your [*sic*] not in position your [*sic*] not going to get a kill".[19] By their estimation, the A-1CM at best only helped with the last 10 percent of the mission, and many thought the new gunsight actually degraded their ability to successfully accomplish the other 90 percent. Within a year, scarcely half of the A-1CM gunsights in Korea still functioned.[20]

The automatic gunsight unsettled the fighter pilot ranks for another reason. Since WWI, the universal measure of a fighter pilot's skill was his record in air combat, with five aerial kills earning the victor the title of *ace*.[21] But if downing a foe was now to be largely automated, what role was then left for the human? The pilots' concerns were reflected in a cartoon published in the 1951 edition of their professional journal, *Fighter Gunnery*. In it, a bloodied fighter pilot lies crushed beneath an enormous box labeled "Gun Sight".[22] One pilot-friendly commentator explained after war's end, "In our enthusiasm for the electronic gadgets which take the thrill out of aerial gunnery and grab the controls from the pilot, we are inclined to forget that science has yet to produce leadership by remote control".[23]

Little more than a year after the A-1CM was thrust into the violent skies over Korea, scores of F-86 pilots were ready to forsake the one tool explicitly designed to make them more lethal. In late summer 1952, 14 Korean War aces approached the Chief of Staff of the Air Force (CSAF) and beseeched him to suspend all development of radar-assisted gunsights. The aces were convinced the "intricate, highly complicated electronic equipment" did not work.[24] One of them had reportedly once remarked he was better

17 E.M. Olsen, "Evaluation of Results of the 1950 USAF gunnery meet", 13 April 1950, CSDL-HC.

18 Evans, *Sabre jets over Korea: a firsthand account*, p.181.

19 Col. R.N. Baker, *Report on F-86 operations in Korea*, 1 April 1953, AFHRA, p.11.

20 Col. F.S. Gabreski, *(End of Tour) Report by Colonel Francis S. Gabreski*, 8 July 1952, AFHRA, p.6; *FEAF Report on the Korean War,* Vol.2, 1954, AFHRA, p.11.

21 R.A. Wohl, *A passion for wings: aviation and the Western imagination, 1908-1918,* New Haven, Yale University Press, 1994, pp.203, 304, n.2.

22 "Birth of a gunsight", *Fighter Gunnery*, Vol.1, No.3, 1951, pp.8-9.

23 H.C. Stuart, "Salute to combat leaders", *Air Force,* Vol.35, No.9, September 1952, p.24.

24 *Accelerated comparison test of the K-14 sight and J-2 fire control system in the F-86E for fighter to fighter combat*, Project APG/ADB/59-A, Eglin AFB, Air Proving Ground Command, 26 September 1952, DTIC, Inclosure 1, pp.5-6;

off using "a piece of chewing gum in the windshield".[25] Summarizing their rationale, the CSAF agreed it might be "most unwise to burden our fighters with 205 extra pounds of gunsight complexity and unreliability".[26] Within weeks, the CSAF ordered an immediate test comparing the performance of the new radar-ranged gunsights with that of their older, manually-ranged cousins. He invited several of the same 14 aces, ranging from lieutenants to colonels, to participate in the test.[27]

The aces were a powerful constituency, but they were not the only one. By the time the CSAF's test kicked off in September 1952, other fighter pilots still flying in Korea had already begun to forge a more collaborative relationship with their gunsight. This fresh crop of pilots opted to discard the strict division of labor originally articulated by the engineers. The pilots began to seek out and study technical manuals that detailed the assumptions built into the gunsight.[28] They scavenged parts and constructed a simulator to further their understanding of the gunsight's operation.[29] And they cultivated collaborative relationships with the engineers, culminating in an innovative radar range-limiter that offered the pilots more control over their automated equipment.[30] Simultaneously, the USAF launched a massive effort to repair and rehabilitate the malfunctioning gunsights.[31] Based on this budding partnership, the pilots still fighting in Korea offered a different assessment, writing the CSAF that "removing the A1CM gunsight" would be a step "made backward instead of forward".[32]

Ultimately, the USAF elected to continue developing its radar-ranged, automated fire

---

FEAF Report on the Korean War, pp.11-12.

25 Quoted in K.P. Werrell, *Sabres over MiG Alley: the F-86 and the battle for air superiority in Korea,* Annapolis, Naval Institute Press, 2005, p.30.

26 Quoted in CG FEAF to CG AF FIVE, "Personal from Smart to Barcus", *History of the 5th Air Force*, Vol.3, 1952, AFHRA.

27 *Accelerated comparison test*, pp.1-2, Inclosure 1, pp.5-6; *FEAF Report on the Korean War,* pp.11-12.

28 *History of the 4th fighter-interceptor wing, Oct. 1951*, AFHRA, chapter 6; *History of the 51st fighter interceptor wing, Feb 1952,* AFHRA, chap.2; *FEAF Report on the Korean War,* pp.11-12.

29 "Radar and gunsight shop", *History of the 4th fighter-interceptor wing, Jul-Dec 1952,* AFHRA, pp.70-71; "Maintenance, supply, communications, and armament history", *History of the 4th fighter-interceptor group*, *Jul 1952*, AFHRA.

30 J.L. Jenkins, "Some notes on the behavior of computing gunsights", *Fighter Weapons Newsletter*, No.1, February 1956, Muir S. Fairchild Research Information Center (MSFRIC), pp.51-53; *FEAF Report on the Korean War*, p.11; *History of the 4th fighter-interceptor wing, Jul-Dec 1952*, p.70; *History of the 6400th air depot wing*, *Feb-Jun 1952,* Vol.1, AFHRA, pp.230-31.

31 *FEAF Report on the Korean War*, p.11; *History of the 6400th air depot wing, Feb-Jun 1952*, pp.114, 261.

32 Brig. Gen. D.D. Hale, USAF, to Commanding General, Far East Air Forces, "A1CM Gunsight", 17 September 1952, in *History of the 5th Air Force, Jul-Dec 1952*, AFHRA, pp.2-4.

control systems, confident that the more advanced systems would be needed in future conflicts. The results of the CSAF-directed comparison test offered hope, concluding that, despite what the aces had said, using a radar-ranged sight doubled the probability of achieving a kill.[33] Still, the USAF wrestled with the realization that aiming and squeezing the trigger *was* only 10 percent of the final solution, and its automated A-1CM could do little to assist the pilot with the other 90 percent. But what if there were even more advanced weapons that didn't rely on a pilot's sharp eyesight, deft maneuvering, and steady aim? Even better, what if these new weapons could actually maneuver in mid-air, on their own, to chase down a wily foe?

The F-4C Phantom II

The USAF's next generation fighter, the F-4C Phantom II, embodied that vision of autonomous weaponry. Originally built for the US Navy but later acquired by the USAF, the F-4C did not even come equipped with an internal gun to shoot down the enemy; it relied solely on a mix of heat-seeking and radar-guided missiles.[34] The engineers were ecstatic. By decoupling the performance of the weapon from that of the aircraft, the new missiles provided unprecedented design flexibility.[35] The new guided missiles could travel higher, further, and faster. Plus, they would be "powerful enough to insure [*sic*] a kill".[36] The pilots were enthralled, too. Describing the allure of the heat-seeking Sidewinder, a USAF pilot remarked in 1958, "the elusive MiG flying 10,000 to 15,000 feet above us in Korea would be real meat for the GAR-8 [Sidewinder]. We might have to revise our standards for qualifying as an ACE in future conflicts".[37]

Almost a decade after that ringing endorsement, and nearly two years after the United States initiated its prolonged bombing of North Vietnam, a group of Phantoms launched from their base in Thailand on a special mission. Led by their famed commander, Colonel Robin Olds, the F-4 fighters successfully goaded 14 North Vietnamese MiG-21s into the air on 2 January 1967. In the ensuing melee, Olds launched two radar-guided Sparrow

33 *Accelerated comparison test*, pp.1-2.

34 *F-4C system package program,* 62 ASZM-52, Air Force Systems Command, rev. September 1963, Air Force Materiel Command/History Office (AFMC/HO), sect.6 and sect.12, p.8; *History of the Tactical Air Command, Jul-Dec 1961*, Vol.1, AFHRA, pp.194-202.

35 G.E. Bugos, *Engineering the F-4 Phantom II: parts into systems,* Annapolis, Naval Institute Press, 1996, pp.27-28.

36 *Air-to-air guided missiles,* Lecture Manuscript, R-49, Tyndall AFB, Air Tactical School, Air University, 16 April 1948, AFHRA, pp.1-2.

37 Capt. R.M. Thor, USAF, "GAR-8", *Fighter Weapons Newsletter,* June 1958, MSFRIC, p.30.

missiles to no avail. Then he launched a Sidewinder, but it too careened off-course shortly after launch. After some hard maneuvering, Olds spied the glint of yet another MiG-21 in the distance. He zoomed high above the horizon, did a barrel roll around the unsuspecting foe, and dove toward its six o'clock. From approximately 4,500 feet away and 15 degrees off the bandit's tail, he fired his fourth missile – another Sidewinder.[38] According to his post-mission report, "Suddenly the MiG-21 erupted in a brilliant flash of orange flame".[39]

All told, Olds and his teammates attempted to fire 31 missiles that day: 19 radar-guided and 12 heat-seeking.[40] For that, they downed a total of seven enemy MiG-21s, and their mission – Operation BOLO – became the USAF's greatest aerial victory of the Vietnam War. The missiles' lackluster performance that day was not atypical, however. It was above average; Olds's team had spent weeks "peaking" their aircraft and weapons for the big mission.[41] Indeed, by the end of the Vietnam War, less than one third of the USAF's guided missiles launched against enemy fighters successfully found their target.[42] The woeful missiles, and the Phantom's singular reliance on them, became a significant concern for the fighter pilots during Vietnam. Analogizing their fate, the pilots explained they had been given "a ten-foot spear" and then locked in a "five-by-five closet" with an enemy armed with "a knife".[43] As in Korea before, the pilots felt victimized by the Pentagon's appetite for complex, automated technologies.[44]

There were other challenges, too, that accompanied the new autonomous weapons. Ever

38 Brig. Gen. R. Olds, USAF, "Briefing to the staff of project CORONA HARVEST on operation BOLO", 29 September 1969, AFHRA; Lt. Col. C.H. Asay, USAF *et al.*, *History of operations BOLO*, in *History of the 8th tactical fighter wing, Jan-Jun 1967*, Vol.2, AFHRA; Maj. G.Y.W. Ow, USAF, *Mission Bolo, 2 January 1967, F-4C vs MiG 21's,* Working Paper 67/3, Directorate, Tactical Air Analysis Center, Headquarters, 7th Air Force, 13 February 1967, AFHRA; J.S. Attinello, *Air-to-air encounters in Southeast Asia: account of F-4 and F-8 events prior to 1 March 1967*, Report R-123, Vol.1, Institute for Defense Analyses, Systems Evaluation Division, October 1967, DTIC, pp.415-46.

39 Quoted in Asay *et al.*, *History of Operations BOLO*, p.41.

40 *Ibid.*, pp.17-21. Maj. Ow recorded 20 radar-guided missiles attempted in *Mission Bolo,* 3.

41 Olds, "Briefing to the staff of Project CORONA HARVEST on operation BOLO", pp.12-14; Asay *et al.*, *History of operations BOLO*, Annex F.

42 M.L. Michel III, *Clashes: air combat over North Vietnam, 1965-1972,* Annapolis, Naval Institute Press, 1997, pp.286-87.

43 Brig. Gen. J.W. Cook, USAFR (ret.), *Once a fighter pilot...* New York, McGraw-Hill, 2002, p.210; Maj. Gen. F.C. Blesse, USAF, Corona Ace Oral History Interview, by Lt. Col. G.F. Nelson, USAF, 14 February 1977, AFHRA, pp.59-60.

44 For a sampling, see Maj. Gen. J.J. Burns, USAF, Oral History Interview, by J. Neufeld, 22 March 1973, AFHRA, pp.3, 9, 14-18; Maj. Gen. R.C. Catledge, USAF (ret.), Oral History Interview, by 1Lt. W.D. Perry, USAF, 30 September 1987 and 9 December 1987, AFHRA, pp.30-32; Blesse, Corona Ace Oral History Interview, pp.59-61.

since Roland Garros opted in 1915 to risk shooting off his own propeller rather than lug around a second person, fighter pilots had reveled in the independence and autonomy of a single seat fighter.[45] The Phantom upended that proud history. The increased automation of the F-4C and its weapons required a second crewmember be reinserted into the cockpit.[46] The front-seater according to the flight manual was the Aircraft Commander; he focused on flying the powerful fighter. It was the back-seater's responsibility to operate the aircraft's complex radar; the flight manual referred to him as the Pilot Systems Operator, but more often he was simply called the GIB – the Guy-In-Back.[47]

Neither individual wanted to share the cockpit with the other. Many front-seaters had already flown single-seat fighters, and the sudden intrusion of a young lieutenant GIB telling him where or how to fly was perceived by some as a threat to their "manhood".[48] The front-seaters felt they "didn't need anybody back there", one GIB recalled.[49] The back-seaters for the most part reciprocated the sentiment. For several years, the USAF GIBs were fully trained pilots usually plucked fresh from flight school.[50] "We had soloed [...] We had flown formation; we had done all sorts of things [...] We felt like young tigers and now we were going to be put in the back seat with another guy", one GIB lamented.[51] Describing their plight, another acknowledged, "[i]t was demoralizing".[52]

Despite the tension, teamwork was essential. The pilot in the front seat might have had a large radar screen mounted prominently at eye level, but he had zero control over its

45 Wohl, *Passion for wings*, pp.203-10; J. Werner, *Knight of Germany: Oswald Boelcke, German Ace*, translated by C.W. Sykes, Philadelphia, Casemate, 2009, p.135; J. Salter, *The hunters,* New York, Vintage, 1956, p.193.

46 "The advantage is here: why the Phantom II has a two man crew", *Air Force & Space Digest,* June 1962; Bugos, *Engineering the F-4*, pp.26-27, 98.

47 *T.O. 1F-4C-1*, USAF Series F-4C and F-4D Aircraft, Flight Manual, 15 February 1979, NMUSAF; *F-4 Aircrew operational procedures*, TACM 55-4, Vol.1, PACAFM 55-5, Vol.1, USAFEM 55-4, Vol.1, Langley AFB, Headquarters, Tactical Air Command, May 1968, AFHRA, pp.1, 29; Col. H.J. Hill, USAF (ret.), Oral History Interview, by J.C. Hasdorff, 12 July 1991, AFHRA, pp.20-21.

48 Gen. L.D. Welch, USAF (ret.), interview by author, 1 October 2013; A.R. Scholin, "New phlyers for the Phantom", *Air Force & Space Digest,* January 1968, pp.46-48; *History of the Tactical Air Command, Jul-Dec 1962*, Vol.1, AFHRA, pp.182-83.

49 Capt. J.L. Hendrickson, USAF, Oral History Interview, by Maj. L.R. Officer, USAF, and H.N. Ahmann, 31 January 1973, AFHRA, pp.61-62.

50 Lt. Gen. J.J. Burns, USAF (ret.), Oral History Interview, by H.N. Ahmann, 5-8 June 1984, AFHRA, pp.215-17; Col. C.R. Anderegg, USAF (ret.), *Sierra hotel: flying Air Force fighters in the decade after Vietnam,* Washington, DC, Air Force History and Museums Program, 2001, pp.40-41; Hendrickson, Oral History Interview, p.99; Lt. Gen. J.T. Robbins, USAF (ret.), Oral History Interview, by J.C. Hasdorff, 24-25 July 1984, AFHRA, pp.105-8.

51 Hill, Oral History Interview, pp.21-22.

52 Brig. Gen. F.K. Everest, USAF (ret.), Oral History Interview, by H.N. Ahmann, 23 September 1988, AFHRA, p.13.

operation. Only the GIB could turn on the radar, and only he could configure its operating mode and adjust its search pattern. Thus, it was the GIB who was responsible for staring at a screen full of "fuzzy blips" and distinguishing between real targets and false ones. When he found a real target, it was the GIB alone who could manually lock on to it. And without that all-important, properly-configured radar lock, the four radar-guided Sparrow missiles slung beneath the Phantom's belly were mere ballast.[53]

When the missiles did work, either radar-guided or heat-seeking, there was still ample opportunity for friction. As mentioned above, aerial victories had always been the ultimate scorecard for fighter pilots, and achieving a kill was recognized as a very intimate, personal affair that relied on individual pilot's cunning and skill. But in the Phantom, who deserved the credit for a kill? The front-seater, for maneuvering the aircraft and pushing the pickle-button to fire the missile? The back-seater, for getting the radar lock? Or the missile, for successfully steering itself toward the target? Automation had entangled the different human and machine contributions and it became impossible to disaggregate them. It took a CSAF decision in spring 1972, more than five years after Olds's famed sortie, to resolve the squabble.[54] But the USAF's decision to recognize both front-seater and GIB each with a full victory credit for every kill was not universally praised. One pilot remarked, "I think it confused the hell out of things".[55]

In fighter aircraft prior to the F-4, such as the F-86, the task of flying the plane was roughly synonymous with employing it in battle – do the first well and the second naturally followed. The complex human-machine relationships embedded within the Phantom altered this dynamic. Pilots became "systems operators".[56] The automation in the Phantom had changed what a fighter pilot did in the air, who he worked with, and how he accomplished his mission. But it had not eliminated human skill – the human operators still played an essential role. A photograph that accompanied the *New York Times* story detailing the USAF's first aerial victory in Vietnam subtly captured the evolving relationship. In the

53 *F-4 aircrew operational procedures*, 29; *F-4C air combat tactics*, Course No.111509F, Phase Manual, MacDill AFB, July 1966, in *History of the 15th tactical fighter wing, Jul-Dec 1966*, Vol.2, AFHRA, pp.18, 21; *T.O. 1F-4C-34-1-1*, USAF Series F-4C Aircraft, Aircrew Weapons Delivery Manual (Nonnuclear), 1 February 1976, NMUSAF, sect.1, p.30; *8th tactical fighter wing tactical doctrine*, March 1967, in *History of the 8th Tactical Fighter Wing, Jan-Jun 1967*, Vol. 2, pp.81-84.

54 R.F. Futrell *et al.*, *Aces and aerial victories: the United States Air Force in Southeast Asia, 1965-1973,* Washington, DC, Office of Air Force History, 1976, p.vi; Col. C.R. Anderegg, USAF (ret.), interview by author, 2 October 2013.

55 Burns, Oral History Interview, June 5-8, 1984, p.219.

56 Welch, interview. See also Olds's similar assessment in "'Ace' fighter wing commanded by 'Aces'", *USAF News Release*, November 1966, AFHRA.

past, a victorious fighter pilot was usually photographed climbing out from his cockpit, alone; now the victory photo was a team shot that included the front-seat pilot, his GIB, and the missile.[57]

The F-15 Eagle

The F-15 Eagle, which took flight in the mid-1970s, was designed to be the "Fighter Pilot's Fighter".[58] With its sleek lines, large bubble canopy, and an internal gun, the Eagle was more reminiscent of the F-86 Sabre than the F-4 Phantom. The USAF longed for the same 10:1 kill ratios that had marked the former's reign of the skies over Korea, too.[59] But the Eagle also incorporated a large, powerful radar and guided missile ordnance, the same equipment that had necessitated a second seat in the Phantom. Nevertheless, fighter pilots were eager to discard the ignominy of flying with a GIB, and consequently the F-15A Eagle would not come with a second seat. In the words of an early F-15 concept paper, "careful attention to automation design" would therefore be necessary.[60]

Some of the automation in the F-15 focused on reducing the pilot's tasks associated with flying the massive, twin-engine fighter jet. For example, an air data computer automatically calibrated the position of the hydraulically actuated, variable-geometry engine inlets based on air temperature and airspeed to ensure proper airflow to the engines.[61] The classic hydromechanical flight control system received an upgrade, too, now boosted by a progenitor of the fly-by-wire control systems used on later fighter aircraft.[62] Not all appreciated the new systems, though. For some prospective pilots, the flight control augmentation system sounded like "black magic" or "something out of *Star Trek"* and they

57 "Pilots describe downing of MIG's", *New York Times,* 12 July 1965.

58 Capt R.J. Hoag, USAF, "Superfighter", *Fighter Weapons Review*, Summer 1974, MSFRIC, p.21; "F-15: a fighter pilot's fighter", McDonnell Douglas advertisement in *Air Force*, August 1971 and October 1971.

59 Gen. J.C. Meyer, USAF, "Air Superiority into the 1980s", *Air Force*, August 1971, p.43.

60 Lt. Col. O'Donohue, USAF, "Development concept paper: new Air Force tactical counter-air fighter (F-X)", 15 September 1968, in *History of the F-15 Eagle*, Vol.2, AFHRA, p.7; D.M. Gillespie, "Mission emphasis and the determination of needs for new weapon systems", Ph.D. dissertation, MIT, 2009, pp.224-25; Brig. Gen. R. Olds, USAF, "The lessons of Clobber College", *Grumman Horizons* 8, No.1, 1968, pp.18-22; Capt. R.S. Ritchie, USAF, Oral History Interview, by Maj. L.R. Officer, USAF, and H.N. Ahmann, 11 and 30 October 1972, AFHRA, p.29.

61 *F-15 Eagle,* Report H446, McDonnell Aircraft Company, 15 July 1969, AFMC/HO, pp.1-2; *USAF F-15 Fighter summary,* Wright-Patterson AFB, Deputy for F-15/JEPO, Headquarters for ASD, 1 May 1971, AFMC/HO, p.12.

62 *T.O. 1F-15A-1*, F-15A/B/C/D Aircraft, Flight Manual, 15 January 1984, NMUSAF, sect.1, p.30; *F-15 Eagle*, p.7; *USAF F-15 Fighter Summary*, p.7.

expressed their concern, "What happens when some berserk electron decides to go on a rampage?"[63] Still, most were enthusiastic. General William Momyer, a three-war fighter pilot veteran and then-commander of the USAF's Tactical Air Command, declared "the F-15 as having 'more potential than a pilot can psychologically take'".[64] Another said it was like "going from an F-150 pickup truck to a Corvette".[65]

The greatest advances in automation, however, centered on the pilot's interaction with the Eagle's radar and weapons. In the Phantom, a human GIB was needed to operate the complex radar. In the Eagle, the radar was specially designed for "one-man operation". The tasks of adjusting the gain and distinguishing between real and false targets was now assigned to the radar itself through new waveforms and automated signal processing.[66] The pilot's interactions with his aircraft weapons systems were also transformed. Instead of relying on a human-GIB intermediary, the F-15 pilot now communicated directly with his aircraft weapons systems using a "very intelligent" 48.5-pound box mounted underneath his feet – the aircraft's central computer.[67]

No longer would the pilot need to verbalize his desired radar or weapons mode changes and wait for the GIB to respond. Instead, he used his HOTAS (Hands-On-Throttles-And-Stick) functions, all monitored and processed by the central computer, to do it "himself". The maze of HOTAS functionality available with the eight separate controls mounted on the throttles and the additional five buttons affixed to the control stick, all of which needed to be actuated in separate, precise sequences, could be daunting. "It is true the throttles are designed for busy hands", one of the USAF's chief F-15 test pilots acknowledged. But, he continued, "the F-15 cockpit design and pilot procedures are simple and uncomplicated and require no more than average physical dexterity and common sense".[68] Pilots abruptly realized that operating the HOTAS correctly was a perishable skill that demanded constant honing. It became known as "playing the piccolo".[69] Mastering its use, the F-15 pilots were

63 Hoag, "Superfighter", pp.22-23.

64 Quoted in "F-15 Rolls Out at St. Louis", *Flight international*, 6 July 1972, p.11.

65 Quoted in Anderegg, *Sierra hotel,* p.150.

66 *Eyes of the eagle: the Hughes APG-63 Radar,* Culver City, Radar Avionics-Aerospace Group, Hughes Aircraft Company, 1976, NMUSAF, p.2; *F-15 avionics and armament mission requirements analysis: report of the ad hoc committee*, Andrews AFB, MD, Headquarters, Air Force Systems Command, 17 August 1970, AFMC/HO, pp.125-26; Hoag, "Superfighter", p.22.

67 Hoag, "Superfighter", pp.26-27; Anderegg, *Sierra hotel*, p.154; Maj. W.G. Flood, USAF, and Maj. J.S. Rodero, USAF, *F-15 Initial operational test and evaluation - air-to-air (Annex A),* TAC Project 71C-223W, Final Report (Interim), Nellis AFB, USAF Tactical Fighter Weapons Center, April 1975, AFMC/HO, p. A:2; *T.O. 1F-15A-1*, sect.1, p.76.

68 Hoag, "Superfighter", pp.22, 25.

69 Anderegg, *Sierra hotel*, p.157.

told, would give them "a significant advantage in the dogfight scenario".[70]

The Eagle might have been custom-built for dogfights, but rapid advances in adversary weapons soon rendered the close-quarters engagement especially lethal for all involved. An F-15 would still likely get the kill, but it was just as likely to be killed in the process.[71] The impressive automation of the Eagle's radar and weapons systems, coupled with continued improvements in US air-to-air missile performance, allowed the pilots to adjust their tactics. Dogfighting promptly gave way to "sorting". Before the transition, targeting enemy aircraft at long-range was largely driven by chance, often resulting in some targets being shot multiple times by different F-15s while other targets emerged from the missile volley unscathed. To avoid this outcome, an F-15 flight leader could "sort" his accompanying fighters to the different enemy targets. However, the tactic required all aircraft in the formation to possess a common radar picture so that each individual pilot could understand and then execute the communicated targeting scheme.[72] Even with two humans in each cockpit, this wasn't possible in the Phantom. But with the Eagle's advanced, automated radar and HOTAS, it suddenly became viable, even if neither function was originally designed with sorting in mind.[73]

The resulting transition in tactics shattered the dogma that since the end of WWI had governed the types of fighter formations flown and the individual responsibilities within those formations. In the past, the formation leader focused almost exclusively on offensive employment while his wingman defended him.[74] Now, with "sorting" all the fighters in a formation had an acknowledged offensive role and all were mutually responsible for ensuring the formation's survival and success.[75]

Automation had changed fighter aviation again. As improvements in aircraft design,

70 *F-15 Eagle Fact Sheet,* 73-2, Washington, DC, Secretary of the Air Force, Office of Information, Internal Information Division, February 1973, AFMC/HO, p.2.

71 Maj. Gen. F.C. Blesse, USAF (ret.), "The changing world of air combat", *Air Force*, October 1977, p.34; "No-win war at Dogbone Lake", *US News & World Report*, 9 January 1978, pp.56-57.

72 Anderegg, *Sierra hotel*, p.160; Col. T. Sokol, USAF (ret.), interview by author, 4 August 2013; Col. C.R. Anderegg, USAF (ret.), interview by author, 2 October 2013.

73 Anderegg, *Sierra hotel*, p.156. Maj. Gen. J. Cliver, USAF (ret.), interview by author, 3 August 2012; Sokol, interview.

74 J. Stocker, "A wing man has to have eyes in the back of his head", *Air Force*, March 1955, p.32; *4th Fighter-interceptor group tactical doctrine*, 22 July 1951, in *History of the 4th fighter-interceptor group, Jul-Sep 1951,* AFHRA, p.9; Maj. Gen. E.A. Bedke, USAF (ret.), Oral History Interview, by D.G. Lamb, 25-26 January 1988, AFHRA, p.41-42; Anderegg, *Sierra hotel*, p.21; Michel, *Clashes*, pp.170-72.

75 Anderegg, *Sierra hotel*, p.163; Flood and Rodero, *F-15 Initial operational test and evaluation – air-to-air*, pp.A:25, A:85.

engines, and advanced flight controls leveled the playing field, the significance of a pilot's stick-and-rudder flying skills waned. In its place, fighter pilots had to develop a new air combat intuition that encompassed their interactions with their radar and central computer, as well as their mutually-supportive flightmates. In the Eagle, fighter pilots were transformed from being "systems operator[s]" into "information integrator[s]".[76] It wasn't the "best hands" that defined a good fighter pilot anymore, one former Phantom and Eagle pilot explained. Now it was who had the "best head".[77]

Beyond the F-15 Eagle

But what if even these human information-integration tasks could be automated? They were in the F-22 Raptor. Relying on a design philosophy known as sensor fusion, the F-22 Raptor was built to automatically gather and process information from a variety of on-board and off-board sensors and fuse it into a single, comprehensive picture of the battlespace.[78] The net result, according to industry publications, was a next-generation fighter plane that promised to "significantly reduce the pilot workload during battle conditions".[79] The pilots had other designs. Rather than merely allow the automation to simplify their tasks, the Raptor pilots instead used the automation to perform new tasks and new missions that until then had been judged nearly-suicidal.[80]

Another alternative to alleviating pilot workload would be to automate the flying itself, removing the human from the aircraft. The Predator and its later cousin, the MQ-9 Reaper, have come a long way from their inauspicious Unmanned Aerial Vehicle (UAV) beginnings. When first unveiled in the mid-1990s, the unarmed RQ-1 Predator was designed to drone along pre-programmed routes collecting intelligence. The two-person human crew was there simply to babysit the aircraft; outside of takeoffs and landings, they would intervene only if the aircraft alerted them to a problem. The first Ground Control Station (GCS) reflected this strict division of labor – just four computer monitors were deemed sufficient

76 Welch, interview.

77 Anderegg, *Sierra hotel*, p.164.

78 D.A. Fulghum and M.J. Fabey, "F-22: unseen and lethal", *Aviation Week & Space Technology*, 8 January 2007; K.W. Greeley and R.J. Schwartz, "F-22 cockpit avionics: a systems integration success story", in *Proceedings of the SPIE* 4022, Cockpit displays VII: displays for defense applications, pp.52-62; "Holding four aces: speed, stealth, agility, and revolutionary avionics", *Avionics Magazine: Special Report, Air Dominance with the F-22 Raptor*, 2002, pp.3-5.

79 R.W. Brower, "Lockheed F-22 Raptor", in C.R. Spitzer (ed.), *The Avionics Handbook*, chapter 32, New York, CRC Press, 2001, sect.5.

80 Fulghum and Fabey, "F-22".

for each crewmember to perform their limited supervisory tasks.[81]

However, once freed from the demands of "flying" their aircraft and with their "cockpits" no longer constrained by the physics of flight, the crews transformed their tasks and their roles. Furnished with an unblinking vantage over a turbulent battlefield, the crews realized they were "uniquely empowered to manipulate, coordinate, and integrate information", one researcher observed, "from an ever-expanding constellation of people and automated tools". Collaborating with a host of others spread across the world, the aircrews worked tirelessly to investigate and "identify patterns and discontinuities" that no machine algorithm could yet resolve.[82] As these tasks evolved, so did the GCS. The number of computer screens eventually doubled and occasionally whiteboards were strung along the sides of the "cockpit" to help the crews manage the information deluge.[83] The later addition of weapons to the aircraft only hastened the human crew's metamorphosis.

In the midst of these changes, the USAF adjusted the name of the aircraft from UAV to RPA (Remotely Piloted Aircraft) to emphasize the humans' role operating it.[84] Regrettably, those human contributions often went unacknowledged. The crews may not have been *physically* present over the battlefield, but they were there *virtually*, and with a closer, clearer vantage than all but a select few. Their missions were incredibly complex and deadly serious, yet the RPA crews' actions were frequently dismissed as though they were simply playing "a big video game".[85] Lost in the rhetoric was that, while the RPA pilots might not have been performing classic piloting tasks using a stick and a rudder, the fighter pilots in an F-15 or F-22 were more likely focusing on the same types of information management tasks as their RPA-brethren.

81 Lt. Col. T.M. Cullen, USAF, "The MQ-9 reaper remotely piloted aircraft: human and machines in action", Ph.D. dissertation, MIT, 2011, pp.64, 254.

82 *Ibid.*, pp.43, 168, 272, 277, 287.

83 *Ibid.,* pp.66, 96, 130, 188, 256, 258.

84 M. McCloskey, "Charting a new course: air force sees culture shift as drone mission gains momentum", *Stars and Stripes*, 27 October 2009.

85 Cited in Cullen, "MQ-9 reaper remotely piloted aircraft", p.9. See also Capt. J.O. Chapa, USAF, "Remotely piloted aircraft and war in the public relations domain", *Air & Space Power Journal,* Vol.28, No.5, October 2014, pp.29-46.

Lessons for automation today

Today's defense planners are frequently implored to "think differently" about the next conflict and the tools that will be needed to prevail. AI and autonomy frequently appear near the top of the list.[86] But too often pundits confuse the two. AI is certainly important, but how we use that AI is much more so. Today's latest AI algorithms are but a new tool that can be applied to the old problems of automation.

The history of USAF fighter aviation offers important lessons about the socio-technical challenges that frequently accompany automation. In each new fighter aircraft, additional automation was developed to alleviate traditional, burdensome pilot tasks. In the F-86, it was calculating the range to an enemy target. By the time of the F-4, a radar and autonomously guided missiles promised to eliminate the pilot's requirements to detect his foe visually and then maneuver into a position close behind it. The F-15 sought to reduce the complexity of the F-4 by further automating the tasks that had previously necessitated a GIB. With the F-22, even those information management tasks succumbed to automation. But the automation in each aircraft was always accompanied by new tasks. Sometimes these tasks were necessary to enable the automation, as was the case in the F-4 Phantom. Other times, the humans pursued new tasks that were previously deemed infeasible, like sorting. These shifts in tasks were usually unforeseen and emerged only through an iterative process of negotiating and renegotiating the human-machine relationships.

As new tasks emerged, the human operators had to develop new skillsets. The transition could be disconcerting to those of previous generations who had staked their reputations on performing the most burdensome but now-automated tasks. To them, the automation was deskilling. Even a former CSAF who had witnessed first-hand the evolution in pilot tasks from aircraft like the F-86 through the F-15 remarked that he "would never have been interested in becoming today's fighter pilot" because they were too often relegated to "simply transporting ordnance".[87] But, in retrospect, one of the great ironies throughout this period is that the automated systems the pilots often feared would relegate them to the sidelines more often generated new imperatives for highly-skilled human pilots to remain tightly-integrated within the system.

The fighter pilot's transformation has gone largely undetected in the public sphere. A simple, one-dimensional job description like *fighter pilot* does not lend itself to the nuances

86 C. Brose, "The new revolution in military affairs: war's sci-fi future", *Foreign Affairs,* Vol.98, No.3, May/June 2019.

87 Welch, interview.

of what fighter pilots actually do in the cockpit – hence the importance of developing a cognitive ethnography to help us uncover the critical relationships between the task, the equipment, and the human operator. But fighter pilots also have little incentive to discuss this transformation with outsiders. Since the first silk-scarved aviators took to the skies in fabric-covered aircraft, fighter pilots have been celebrated in print and film as uniquely endowed to slip earth's surly bonds, possessing a cunning wit, cat-like reflexes, and hawk-like vision. Much has changed about flying a fighter jet over the last five decades, but the fact that today's pilots still thunder into the skies provides a powerful and inviolable connection to the intrepid aviators that went before them. Our understanding of how automation affected them, and how it could affect us, suffers because of it.

Thus, revisiting the history of USAF fighter aviation is particularly useful because it illuminates the pervasiveness of what Mindell described as the three myths of autonomy, and it validates his assessment that "automation often changes the type of human involvement required and transforms it but does not eliminate it".[88] Just as they have in the past, these three myths distort our understanding about autonomy in the future, and AI's potential within it. Whether aided by AI or not, the central question remains: what is the goal of the automation? Is it for our machines to replace us or to complement us? These are important questions, especially when we are talking about hypothetical AI-empowered, terrorist-hunting drones.

88 Mindell, *Our Robots, Ourselves*, p.10.

4

Intelligent machines and the growing importance of ethics

Andrea Gilli, Massimo Pellegrino and Richard Kelly

Thanks to their superior computational power and access to broader data, intelligent machines are increasingly outperforming human beings in prediction. As Agrawal, Gans and Goldfarb remind us in their chapter, however, prediction is different from decision as decision entails also philosophical, ethical, moral, legal and cultural considerations. An algorithm may determine a cancer patient's survival chances after chemotherapy, but this important information must still be translated into a decision: whether such chances are low or high, should the patient be treated at all? Should the patient be given priority *vis-à-vis* other cancer patients with the same or different survival chances (because of gender, race, wealth or other reasons?), or even should cancer be brought up or down in hospitals' priorities? Prediction helps make decisions, but decisions do not just require prediction.

A recurrent theme of this monograph is that as machines take over tasks from human beings, human beings specialise in more important activities. This chapter goes one step further: in the age of artificial intelligence, a fundamentally human domain, that of ethics, acquires more salience.[1] Yeung and Celaya as well as Fino, in their respective chapters, provide different evidence and analytical frameworks to appreciate these issues. Agrawal, Gans and Goldfarb note that machines are *per se* unfit for striking payoffs – they have little judgement. Human beings must then step in and provide it: when we entrust decisions to intelligent machines, we need to clarify, for them, our philosophical, ethical, moral and legal preferences. There is a related consideration. Machines do what coders write on the basis of the data they have access to. Such codes and data implicitly or explicitly reflect broader, and potentially contentious, philosophical, ethical, moral, cultural and legal stances. As decisions are automated, we must make sure that as human beings we are fully aware and in

1 B.D. Mittlestadt *et al.*, "The ethics of algorithms: mapping the debate", *Big Data & Society*, 2016, pp.1-21.

control of the assumptions and implications of such codes and data.[2]

The current chapter elaborates on these issues. It first reports the increasing attention devoted to ethics. Then it discusses the reasons behind this development. Next, it looks in detail at the practical questions that are calling for answers. After a brief summary of the main ethical principles on which the global discussion hinges, the chapter highlights their challenges and trade-offs. It concludes with some thoughts and implications for defense and security.

Growing attention to ethics

Over the past few years, attention to the interaction between artificial intelligence and ethics has grown substantially.[3] As a result, many actors have stepped into this important debate: from the OECD[4] to the European Union[5] and UNESCO,[6] from private companies like Microsoft,[7] IBM,[8] Google[9] or Accenture and PwC[10] to industry associations – among many others.[11] One of the most interesting contributions comes from the Institute of Electrical and Electronics Engineers Global Initiative on Ethics of Autonomous and Intelligent Systems that has comprehensively mapped the issues and challenges that are emerging or will in the near future.[12]

A key question, at this stage, is why these different actors, from academic scholars to

2 L. Floridi *et al.*, "AI4People – An ethical framework for a good AI society: opportunities, risks, principles, and recommendations", *Minds and Machines*, Vol.8, No.4, 2018, pp 689-707.

3 For a summary, see A. Jobin *et al.*, "Artificial Intelligence: the global landscape of ethics guidelines", *Nature Machine Intelligence,* Vol.1, 2019, pp.389-399.

4 OECD, *Recommendation of the Council on Artificial Intelligence*, Paris, OECD, 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed 11 November 2019).

5 European Group on Ethics in Science and New Technologies, *Statement on Artificial Intelligence, Robotics and "Autonomous" Systems*, Brussels, European Commission, 2018.

6 UNESCO, *Beijing consensus on Artificial Intelligence and education*, UNESCO Digital Library, 2019, https://unesdoc.unesco.org/ark:/48223/pf0000368303 (accessed 11 November 2019).

7 G. Shaw, *The future computed: AI & manufacturing*, Seattle, Microsoft, 2019.

8 A. Cutle *et al.*, *Everyday ethics for Artificial Intelligence*, New York, IBM, 2019.

9 Google, *Perspectives on issues in AI governance*, Mountain View, Google, 2019.

10 S.C. Tiell and J. Metcalf, *Accenture, universal princpiles of data ethics: 12 guidelines for developing ethics codes*, Accenture, 2016; A. Rao *et al.*, *A practical giude to responsible Artificial Intelligence (RAI)*, NewYork, PwC, 2019.

11 Association for Computing Machinery US Policy Council, *Statement on algorithmic transparency and accountability*, Washington, 2017.

12 The Institute of Electrical and Electronics Engineers Global initiative on Ethics of Autonomous and Intelligent Systems, *Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems*, New York, IEEE, 2019.

think tanks, from international organisations to NGOs, from private companies to industry association have spent so much time and effort to discuss ethics. At first glance, it seems even paradoxical that – in the age of growing computing power, big data and algorithms – we are spending time to debate abstract issues, and even philosophy. Aren't we told by economists and education institutions that we need more graduates in science, technology, engineering and maths (STEM) as opposed to those in humanities? If, so, why, then, do such STEM people discuss philosophy rather than work on hard technical issues? At closer scrutiny, this development is far from surprising: as intelligent machines take over more and more tasks, their actions will have deeper and broader consequences. Action, however, is informed and assessed by ethical considerations and thus in order to advance, adopt and ultimately improve intelligent machines, it is first necessary to agree and define the ethical perimeters and parameters of their tasks.

Human action and human values

In order to understand such ethical perimeters and parameters of intelligent machines' activities, it is first useful to grasp the material reasons behind this increasing importance allocated to ethics. At the most basic level, the reasons lie primarily with the three main differences between machines and human beings. First, intelligent machines have much more computational power and thus can solve more quickly a higher number of more complex number functions. As Eric Brynjolfsson and Andrew McAfee note, a calculation that would have taken 89 years in 1982 can now be made in a few seconds.[13] Next, because of the improvement and diffusion of digital technologies, intelligent machines have access to many more input data. Digital data are doubling every two years and 90 percent of data available in the world was created in the past 24 months.[14] Last, but not least, intelligent machines are not intelligent in the way human beings are.[15] They do not decide to go to the beach on a workday or that they prefer raspberry to apple.[16] What they do, in fact, strictly depends on what coders have written on the algorithm, intentionally or unintentionally, and on the data they have access to and thus what data, implicitly or explicitly, reflect.

13 E. Brynjolfsson and A. McAfee, *The second machine age: work, progress, and prosperity in a time of brilliant technologies*, W. W. Norton & Company, New York, NY, 2014.

14 D. Reinsel, J. Gantz and J. Rydning, "Data age 2025: the digitization of the world from edge to core", *White Paper*, International Data Corporation, Washington, DC, November 2018.

15 J. Searle, "Minds, brains, and programs", *Behavioral and Brain Sciences,* Vol.3, No.3, 1980, pp.417-424; A. Touring, "Computing machinery and intelligence*", Mind, New Series*, Vol.59, No.236, 1950, pp.433-460.

16 J. Markoff, *Machines of loving grace: the quest for common ground between humans and robots,* New York, ECCO, 2015, p.333.

Even assuming there is such a thing as a perfect algorithm, if input data are deeply biased, predictions may lead to incorrect predictions or ethically questionable decisions. Recurrent problems related to minorities or gender stereotypes in hiring procedures are just one possible example.[17]

This poses serious problems that are further exacerbated by their opacity and the uncertainty in which they operate. The traditional example is that of an imminent car accident involving an autonomous car. The onboard software has to decide whether to veer right, and face 100 percent chances of killing a young girl, or veer left and, for instance, having 50 percent chances of killing an elderly couple. The number of examples is infinite and could include a young promising professional and an unemployed person, or a wealthy sportsman and an underground artist and so forth. The key challenge remains: what should the autonomous car do? Its actions will be driven by the algorithm and the data, but this still poses a question: whether to base the decision on gender, age, probabilities, number of fatalities, their contribution to society or on any other consideration. Human beings face only a fraction of these challenges as they just do not have access to all those data and cannot instantly make accurate probabilistic prediction. Thus, they can only be held responsible for their compliance with the law – not about the ethical implications of their actions. Autonomous systems, in contrast, can. The problems is that, as Agrawal, Gans and Goldfarb remind us, while human beings are good at decision but bad at predictions, machines are good at predictions but not designed to make decision. They cannot determine what is right or wrong. Not only do we have to agree first on these parameters, but we have then to define this for them, and finally we have to make sure that machines will follow instructions. Otherwise, machines' decicions and actions will just reflect what coders, and data, implicitly or explicitly suggest: for instance veer right or left depending what social group already suffers the highest number of car accidents.[18]

Why a focus on ethics is needed

People in uniform know legendary US Air Force Maj. John Boyd and his OODA loop: observe, orient, decide and act.[19] Intelligent machines are basically taking over the two Os, observe (data-gathering) and orient (data-processing). The previous section has highlighted the roadblocks when it comes to the D of decisions. Thus, machines can, at

17 M. Whittaker *et al.*, *AI Now Report 2018*, New York, AI Institute, 2018.

18 A. Etzioni, O. Etzioni, "Incorporating ethics into Artificial Intelligence", *Journal of Ethics*, Vol.21, 2017, pp.403-418.

19 G.T. Hammond, *The mind of war: john boyd and American security*, Washington, DC, Smithsonian Books, 2001.

best, identify optimal strategies – but they cannot say, whether a choice was right or wrong. The challenges are, however, broader. What is the purpose of intelligent machines? Why and how should data be gathered, stored and processed? Who should write algorithms and what guidelines should she follow?[20]

These are important questions that cannot be addressed in this chapter. For the sake of clarity, the following paragraphs try to summarise the debate – and the eventual consensus – that has emerged over the past few years. These are in fact the recurrent principles on which many organizations and institutions tend to converge.

Human-centered AI. What is the purpose of AI? Are there any boundaries to its actions and effects? The first, generally shared, ethical imperative is that AI must be human-centered, i.e. be based on fundamental human rights and thus algorithms, big data and – more in general – their applications should be designed for improving human beings' living conditions, including their dignity and rights, freedom, autonomy, their purpose in life as well as the natural environment surrounding them.[21]

Explainability and transparency. A key challenge regarding the work of intelligent machines is to understand the logic behind their decisions. As a recent document from UNESCO notes, high-functioning "AI such as AlphaGo or Watson can perform impressively without recognizing what it is doing". In fact, "AlphaGo defeated a number of Go masters without even knowing that it was playing a human game called Go".[22] Because of the way certain algorithms, for example deep learning techniques, are written, it is often challenging for human beings to understand how machines reach some conclusions.[23] Most stakeholders agree that the development, deployment and future employment of artificial intelligence require human beings to be able to make sense of machines' choices (explainability) and the way algorithms are written and operate to be clear (transparent) – i.e. we must know what are the fundamental principles around which algorithms are designed. Whether an algorithm identifies a certain object as an appropriate military target or a software classifies a certain X-ray signature as a bone fracture, it is of utmost importance that both *ex ante* and *ex post* we can track how those conclusions were reached. Campaigns about explainable

20 S. Etlinger, *AI ethics: making difficult decisions in corporate settings*, San Francisco, Altimeter, 2018; S. Etlinger, *Ethical AI: a quick-start guide for executives,* San Francisco, Altimeter, 2019; S. Etlinger, *Innovation + trust: the foundation for responsible Artificial Intelligence*, San Francisco, Altimeter, 2019.

21 Independent High-Level Expert Group on Artificial Intelligence, *Building trust in human-centric Artificial Intelligence*, *ethics guidelines for trustworthy AI*, European Commission, 2018.

22 Executive board, *Preliminary study on the technical and legal aspects relating to the desirability of a standard-setting instrument on the ethics of Artificial Intelligence*, Two hundred and sixth session, Paris, UNESCO, 2019, Annex A, p.3.

23 B.D. Mittlestadt *et al.*, *The ethics of algorithms*, pp.1-21.

AI and transparent AI are nothing but a reflection of these concerns in order to avoid unfortunate cases and situations.[24]

Responsibility and accountability. The issues of explainability and transparency lead directly to the next question, namely accountability and responsibility. We often tend to think of machines as inherently superior to human beings. Yeung and Celaya caution that machines' performance depends on their interaction with human beings. Citing David Mindell, Fino in his chapter reminds us of the myths of automation.[25] These insights are useful because errors, malfunctions and mistakes are inevitable. Thus, clear lines of responsibility are necessary to facilitate the adoption of artificial intelligence. Who is responsible, in the case of a car crash, for the fatality? The coders who wrote the algorithm, the data scientists who provided training data, the car manufacturer responsible for the software integrator or the human in the loop? Vulnerability to cyber attacks complicates and at the same time strengthens the need for clear lines regarding accountability.

Reliability and Safety. The need to establish accountability and responsibility as well as to put human beings at the center of the AI transformation directly leads to another requirement, namely ensuring reliability and safety, two additional principles on which intelligent machines must be built and designed – in order to avoid errors, mulfunctioning and sabotage.[26] Countries, companies and other actors may be tempted to rush into deploying autonomous systems to beat the competition or achieve some immediate benefits. This may put human life at risk and, more generally, undermine trust in such platforms with the end result that adoption and diffusion may slow down.[27]

Fairness and inclusion. While hard disciplines like computer sciences are often thought to be less amendable to debate, the increasing use of algorithms opens up some critical questions. For instance, who makes AI and for whom is it made? People "who lack educational opportunities", for instance, may not contribute in writing algorithms and thus their values, principles and stances may be neglected, given the importance coders play, with the end result that such communities are penalized. Similarly, as Kate Crawford from Microsoft and the AI Now Institute explain, "data will always bear the marks of its history". Some countries, groups or individuals, for example, may have less or different

24 M. Brundage *et al.*, *The malicious use of Artificial Intelligence: forecasting, prevention, and mitigation*, Oxford, Future of Humanity Institute, 2018.

25 D.A. Mindell, *Our robots, ourselves: robotics and the myths of autonomy*, New York, Viking, 2015.

26 P. Scharre, *Autonomous weapons and operational risk,* Washington, Center for New American Security, 2016.

27 P. Scharre, "Killer apps: the real dangers of an AI arms race", *Foreign Affairs*, 2019.

digital presence and thus data on which algorithms are based, or may neglect their habits.[28] The end result is reinforcing marginalization and discrimination of certain communities.

Privacy and Data governance. The last, but not least, recurrent principle concerns data privacy and data governance. Why are data gathered? Who is collecting data, and where are they stored? How deep should data-gathering go? For what purposes should data be used? These questions ultimately refer to individual rights and freedoms and the fact that data, the key component of the AI revolution, are becoming easier to gather and to store. However, this does not mean that individual privacy must not be respected and prioritized. The EU's General Data Protection Regulation, for instance, requires explicit consent to allow an organization to gather and store data about individuals. There are, obviously, other ethical requirements about data storage for instance, servers can be hacked. Privacy, however, also calls for anonymisation of data to prevent any malevolent actor retrieving the actual identity of a certain population.

Ethical imperatives and contentious issues

Ethics is not a natural science. It neither aspires to, nor possesses a method, for reaching conclusive evidence. However, ethics matters and the specific approaches human beings adopt in these spheres can exert a strong influence on their personal life, as well as on world affairs. There are three main pillars in Western moral philosophy: Aristotelian virtue ethics, Kantian deontology, and consequentialism – although all other cultures have developed their different formal moral frameworks, from Confucian to Shinto and Hindu thought to Islam and Judaism.[29] For the sake of this discussion, these three ethical approaches locate the source of moral actions at different levels and identify different parameters to assess decisions. Virtue ethics stems from Aristotelian philosophy, and its premise holds that the ultimate goal of human beings consists in a constant improvement of themselves as well as of the people and environments surrounding them. Immanuel Kant's deontological ethics is based on the notions of duty, rationality and universal rules: human beings can rationally create and abide by rules that allow for a duty-based ethics to emerge. Utilitarian ethics, or consequentialism, in contrast, is based on the consequences of one's actions – paying attention, however, that they is not short-termed or in contradiction with some fundamental individual rights.

28 A. Hagerty and I. Rubinov, "Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence", *arXiv*, 2019.

29 The Institute of Electrical and Electronics Engineers Global initiative on Ethics of Autonomous and Intelligent Systems, *Ethically aligned design*, IEEE, pp.37-41.

These three approaches cannot, fundamentally, be reconciled when it comes to foundational issues. In simple terms, each person decides to behave as he/she prefers in every day interaction. Such individual preferences and choices are to be respected. However, when intelligent machines start taking decisions with real-life consequences on other human beings, their ethical preferences acquire new salience.[30] A key question emerges: granted all different ethical perspectives comply with the law, who, and on what basis, should decide what ethical stance should intelligent machines be based on? How often should such stances change and according to what parameters: elections, bureaucrats' decisions, or other factors? To go back to the car accident example, would we rather accept that the young girl, or the elderly couple, is killed – and who decides that?

Some prominent experts even believe that for instructing machines to behave ethically is a false challenge. As long as machines comply with the law, each individual should choose on behalf of their intelligent machines. Even in this case, however, the solution is not bereft of problems. For instance, many individuals just stick to the default options.[31] Similarly, since intelligent machines are based on data, individuals' behaviour may actually lower ethical stances. Last but not least, there is still the question of organizations-owned intelligent machines, from hospitals to security forces.

There are, however, also other challenges. In particular, some of the principles discussed in this chapter are in direct contradiction. The ethical imperative of prioritizing safety is directly at odds with privacy: safety calls for data-gathering while privacy calls for protecting individuals' rights and thus limiting data-collection. Similarly, reliability calls for common standards and thus harmonisation. But harmonisation is potentially at odds with inclusion. This section has just mentioned three main ethical positions in philosophy; however, they are based on Western traditions and thus other cultures may not see themselves represented in such underlying principles and values. Some suggest that these tensions should be addressed through democratic processes. This position is also difficult either because some issues may be very technical or because not all actors – at the local, regional or global level – may see democratic deliberation as the most appropriate method.

Additionally, human-centered AI is based on the premise that intelligent machines should promote human dignity, purpose in life and standards of living. AI can favor innovation, but innovation does not equally favor all individuals. Then, in the case of AI,

30 L. Floridi, "Translating principles into practices of digital ethics: five risks of being unethical", *Philosophy & Technology,* 2019.

31 A. Eztioni and O. Etzioni, "Incorporating ethics into Artificial Intelligence", *Journal of Ethics*, Vol.21, 2017, p.413.

innovation and privacy may be at odds, as data feed algorithms and thus limitations to data-collection for privacy reasons may stifle innovation.[32] The very first grand principle on which our discussion started, human-centered AI, could even be questioned by some who, rightly or wrongly, may believe that other political entities are ontologically more important, including the relevant social group, the nation, or a certain set of beliefs (be it a political ideology or a religion).[33]

This discussion about ethics brings us to a final, controversial issue: while we often hear about an imminent AI arms race, what has been written so far suggests that we may see a different type of competition in the near term. Precisely, countries, actors and organization may start competing to promote their ethical view on AI.[34] The European Commission, for instance, maintains that it will "continue its efforts to bring the Union's approach to the global stage and build a consensus on a human-centric AI".[35] The Pentagon's AI strategy has a subtitle on "leading in military ethics", in which it writes that one goal will consist in "upholding and promoting our Nation's values".[36]

What does it mean for the world of defense?

What does all this mean for the world of defense? The current political debate, especially in Europe, is monopolised by the marginal issue of lethal autonomous weapons (LAWs). Lethal autonomous weapons are not here yet; they will face major technical, tactical and operational challenges but, most importantly, they face the opposition of the strongest actor ever: the very armed forces that are supposed to adopt them. Most generals, admirals and air marshals – at least within NATO – are all too aware of the political risks of employing lethal autonomous weapons and thus shy away from this technology. This does not mean that this technology will not be developed or adopted, but at this stage it is blinding most observers to much more pressing and important challenges and risks – even when it comes to ethics.

---

32 E. Chivot and D. Castro, *The EU needs to reform the GDPR to remain competitive in the algorithmic economy*, Washington, Center for Data Innovation, 2019.

33 UNIDIR, *The weaponization of increasingly autonomous technologies: considering ethics and social values,* UNIDIR Resources, 2015, https://www.unidir.org/publication/weaponization-increasingly-autonomous-technologies-considering-ethics-and-social-values (accessed 12 November 2019)

34 L. Floridi, *Translating principles into practices of digital ethics.*

35 European Commission, "Building trust in human-centric Artificial Intelligence", *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions*, Brussels, 2019, p.8.

36 US Department of Defense, *Summary of the 2018 Department of Defense Artificial Intelligence strategy: harnessing AI to advance our security and prosperity*, Washington, US Department of Defense, 2019, p.15.

This chapter has tried to summarize and highlight the state of the debate, illustrating the technological and political reasons behind this growing attention to ethics. Importantly, while the chapter started by reporting the important initiatives and documents that have been launched and have appeared in the past few years, there are not many comparable efforts in the security world – if we exclude the ongoing talks in Geneva about lethal autonomous weapons and the recent document released by the US Department of Defense's Defense Innovation Board on AI ethics principles.[37] Likely, this will have to change soon. In particular, as AI technologies and platforms will be adopted by armed forces and ministries of defence around the world, such actors will have to clarify their ethical stances and imperatives. For instance, constant, comprehensive access to biometric data about uniformed personnel may optimise deployment, rotation and promotion – just to mention a few. Should, however, people in the armed forces give up their total privacy? And what are the risks involved – including that the adversary can access and potentially compromise such data? Similarly, who should decide how an autonomous logistical convoy will have to behave in case of imminent accident: should it prioritise the life of the military personnel it is supporting, should it protect the colonel vs. the caporal and to what extent should civilian casualties be avoided? International Humanitarian Law, on which the next chapter from Marsan and Hill elaborates, provides a starting point but not a conclusion. The entire debate is, however, flipped: in the age of algorithms, we must decide *ex ante* about challenging situations. In the past, we just had to assess *ex post* if all procedures were correctly followed. This chapter has illustrated some of the most pressing challenges and ethical concerns: increasingly, countries as well as their security partners will have to find some answers.

37 Defense Innovation Board, *AI principles: recommendations on the ethical use of Artificial Intelligence by the Department of Defense*, Washington, US Department of Defense, 2019.

5

International law and military applications of Artificial Intelligence

Nadia Marsan and Steven Hill [1]

As an Alliance of 29 Nations with collective defence as one of its core tasks, NATO must be prepared to address emerging threats and security challenges arising from the development of new technologies. The development and use of artificial intelligence (AI) presents both opportunities and challenges to the North Atlantic security environment. The level of discussions within NATO on the implications of AI for the Alliance's multinational operations has been steadily increasing. In March 2018, Allied Command Transformation (ACT), NATO's adaptation hub located in Virginia in the US, organized an informal workshop with NATO ambassadors and military representatives to discuss the impact of the development of disruptive technologies on the Alliance. One take-away from this discussion – as well as others – is that Nations may wish to discuss some of the legal implications of this emerging technology in a multilateral forum such as NATO.

How Allies respond to this changed security environment, both individually and collectively, raises a number of legal questions. The cross-cutting nature and widespread impact of AI-enabled technologies necessitates multinational discussion and debate. Although an increasing number of States, including some NATO Allies, and the European Union are developing AI strategies,[2] and although there have been extensive discussions at the international level on specific topics such as lethal autonomous weapons systems (LAWS), debates on the legal aspects of the full range of potential military applications of

---

1 This chapter is based on a paper developed by both authors for the NATO Science and Technology Organization's annual scientific conference (IST-160 Specialists' Meeting on Big Data and Artificial Intelligence for Military Decision Making held in Bordeaux, France from 30 May to 1 June 2018). It represents the personal views of the authors and does not necessarily reflect the views of NATO or its Allies. The authors gratefully acknowledge research assistance provided by Elif Ece Ozturk.

2 Including, for example, the French AI strategy "AI for Humanity" released in March 2018; the European Commission's Communication "Artificial Intelligence for Europe" released in April 2018; and, the United States' Executive Order on Maintaining American Leadership in Artificial Intelligence issues, February 2019.

AI are still at the fairly nascent phase. Such discussions are largely dominated by academic writing[3] and there is a dearth of State views. As of yet, there are no NATO policies providing guidance on how to address the development and multinational use of AI technology in a military context.

This chapter provides some preliminary observations from the personal perspective of NATO legal advisers who have begun to encounter these issues, rather than on the basis of agreed NATO positions or doctrine. It is therefore intended to introduce rather than resolve some of the legal issues that are likely to arise in future AI-related discussions within NATO. This chapter begins by presenting the concept of "legal interoperability", one of the tools that legal advisers working in NATO seek to promote. It then introduces some of the current legal issues and debates surrounding the military development and use of AI, including questions pertaining to accountability. The paper argues that given the rapidly evolving technology and the asymmetric approach and capabilities of nations on the matter, efforts within the Alliance should focus on ensuring NATO's "legal preparedness" so that collective action is not thwarted by legal hurdles and mismatched legal approaches.

The NATO context: legal interoperability

NATO is an Alliance of values, and one of the core values that the Alliance defends is the rule of law. The commitment to respect the rule of law is enshrined in the preamble to the 1949 North Atlantic Treaty and is reaffirmed in declarations by NATO Heads of State and Government at their regular Summit meetings. It is a key element of NATO and NATO-led operations and it is necessary to ensuring legitimacy and securing public support. Compliance with international law is an equally essential component of NATO's success in all of its endeavours and activities. Correspondingly, the demand for legal advice from over 70 legal offices that make up the NATO legal community has never been higher. The scope of issues on which the typical NATO legal adviser is called to advise on can be daunting and guidance on Allied views is not always available, especially in an easy-to-access consolidated format. This is especially true in the context of emerging technologies such as the security impact of AI-enabled tools.

NATO Allies each have their sovereign domestic legal systems and are bound by different international legal obligations. Allies also often have their own understanding on what international law obligations are applicable to them and under what conditions. As

3 A. Deeks *et al.*, Machine learning, Artificial Intelligence, and the use of force by States, *Journal of National Security Law & Policy*, Vol.10, 2019.

an alliance of 29 nations, NATO takes all decisions by consensus and the main challenge in such a multilateral environment is enabling Nations to act together in line with the individual legal obligations of each, taking account of the differences in applicable legal parameters. A key technique that legal advisers use to help achieve consensus despite these legal differences is "legal interoperability".

"Legal interoperability" is derived from the military concept of interoperability of forces, whereby the military forces of the 29 Allies are able to work together to achieve common objectives. NATO doctrine defines the interoperability of forces as "[t]he ability of the forces of two or more nations to train, exercise and operate effectively together in the execution of assigned missions and tasks".[4] In the 2016 Warsaw Summit Declaration, Allied Heads of State and Government referred to the need for interoperability to accomplish NATO's goals.[5] The 2018 Brussels Summit Declaration also referred to the need to increase interoperability as an element of increased burden-sharing among Allies.[6]

From a legal perspective, interoperability refers to the need for Nations to work together in a variety of contexts despite the application of differing legal frameworks and obligations. As one observer of NATO operational law defined it, "'[l]egal interoperability' is understood […] as the ability of the forces of two or more nations to operate effectively together in the execution of assigned missions and tasks and with full respect for their legal obligations, notwithstanding the fact that nations concerned have varying legal obligations and varying interpretations of these obligations".[7] Legal interoperability relies on a broad understanding of those areas of legal divergence amongst Allies, which requires careful analysis of the legal positions expressed by the relevant Nations. In the case of new technologies, this understanding is difficult to acquire, in part because Nations have not had the opportunity to set forth their legal views or have refrained from doing so in order to prevent the loss of a potential technological edge.

Within NATO, there is often a diversity of legal views on core issues of international law. Let us take, for example, the laws of war, where 27 of NATO's 29 Allies are party to the Additional Protocols to the Geneva Conventions. As a military Alliance founded on the rule of law, these sometimes different legal obligations can be challenging. The challenges

4 *AAP-06: NATO glossary of terms and definitions*, NATO Standardisation Office, 2014.

5 See *NATO Warsaw Summit Communiqué*, NATO e-Library, 2016, paragraph 49, https://www.nato.int/cps/en/natohq/official_texts_133169.htm (accessed 11 November 2019).

6 See the *NATO Brussels Summit Declaration*, NATO e-Library, 2018, paragraphs 3 and 31, https://www.nato.int/cps/en/natohq/official_texts_156624.htm (accessed 11 November 2019).

7 M. Zwanenburg, "International humanitarian law interoperability in multinational operations", *International Review of the Red Cross*, Vol.95, 2013, pp.681-4.

are evident when there is a need for a rapid decision by Allies based on consensus and requiring legal justification acceptable to all. For these reasons, legal interoperability has proven to be indispensable to rapid decision-making and is facilitated by Allies' pragmatism when faced with questions of collective security, emphasizing and relying on those elements that are common to all rather than on those elements of divergence.[8]

Selected military applications of AI

There are already applications of AI that are of interest to a multilateral security organization such as NATO. The international debate about AI has often focused on LAWS. However, the more likely applications of AI are in far different areas. For the purposes of this paper, it would be useful to highlight three areas where AI has been and is currently being used with effects, both positive and negative, on Allied security.[9] On the positive side, AI could enhance NATO's intelligence, surveillance and reconnaissance (ISR) capabilities as well as its ability to analyse personal data for appropriate purposes. On the negative side, AI could be used against NATO Allies as part of the disinformation campaigns that have come to characterize hybrid warfare.

Intelligence, Surveillance and Reconnaissance

Effective security and military decision making depends on reliable situational awareness. In recent years, technological developments have enabled the autonomous collection of massive amounts of data, including measurements, satellite and infrared imagery, and electronic signals, through ISR systems, such as unmanned aerial vehicles (UAV) and satellites. The ISR process begins with information gathering through sensors to intelligence analysis where the information is brought together towards supporting effective operational decision-making and situational awareness. The increased autonomy of sensors calls for an equivalent autonomy in analysis through systems able to identify patterns and trends in the data.[10] The amount of data is so large that its effective exploitation cannot rely on traditional human analysis for increased situational awareness. However, the systemic and

8 S. Hill and D. Lemétayer, "Legal issues of multinational military operations: an Alliance perspective", *The Military Law and the Law of War Review,* Vol.55, 2017, pp.13-33.

9 The following examples are drawn from: G. Allen and T. Chan, *Artificial Intelligence and national security,* Harvard Kennedy School, Belfer Center Study, 2017, https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf (accessed 12 November 2019).

10 *Ibid.*, p.27.

unsupervised use of AI to process and analyse ever larger sets of data could potentially raise concerns of cognitive bias imported into the analysis – underlining the necessity to keep a human being in the analytical process. Human analysis relies on critical thinking which, so far, has not been perfected in AI-enabled technology.[11]

Analysis of personal data

Combining the data gathered through ISR sensors with personal information gathered from various databases and from social media platforms could enable the automatic analysis of personally identifiable information by machines which could facilitate the identification of threat actors. Within the context of NATO operations, Allies have recognised that capturing and matching biometric samples from individuals encountered in operational theatres could provide a vital capability to support the identification of anonymous operatives to persons, places, items and events, thereby enabling effective, targeted defensive action against threat networks. In the national security and defence field, such calls for greater sharing, while often made, also come up against a wide range of legitimate concerns, including privacy and potential human rights concerns.[12] Between the United States and Europe, for example, Allies have different national laws with respect to the privacy rights affected by the handling of personal information, including different obligations under international law.

Disinformation campaigns

There have been considerable improvements in the production of convincing and high quality virtual reality capabilities.[13] Virtual reality has also become much more accessible in a variety of contexts, for example to optimise museum visits and to support military battlefield training. These technologies can also be used to produce audio and video spoofs, which can be used to manipulate the public information space. Not only is the fundamental value of freedom of speech threatened when there is a proliferation of unreliable narratives, but the proliferation of "fake news" introduces doubt and feeds public debate on the facts underpinning security related events. In such information environments, government and military decision-making is "slowed" by intense scrutiny on questions of fact: "the

11 *Ibid.*

12 NATO's work in this area builds upon *UN Security Council Resolution* 2396, S/RES2396, UNDOCS, 2017, https://undocs.org/en/S/RES/2396(2017) (accessed 12 November 2019).

13 For an excellent overview of this area see G. Allen and T. Chan, *Artificial Intelligence and national security,* Harvard Kennedy School, Belfer Center Study, 2017, note 11.

existence of widespread AI forgery capabilities will erode social trust, as previously reliable evidence becomes highly uncertain".[14] Virtual reality could also be used to interfere with high-level decision-making, by crippling situational awareness with "fake intelligence", as well as with military command and control, through the introduction of fake orders by an adversary. AI-technology is now even being developed to verify content in order to facilitate the identification of fake, doctored material. As in the two previously mentioned areas, the autonomy of such technologies, independent from human control and oversight, raises the greatest challenge.

Accountability issues related to AI-enabled technology

International and domestic law offers some tools to cope with the effects of AI-enabled weapons from a multinational collective defence perspective. New technology does not necessarily require new laws and it is not the aim of this paper to propose the creation of new legal frameworks. The applicable legal framework will depend on the technology used and on the actual and potential effects caused by that technology and can include various national laws, human rights law, the law of state responsibility, international humanitarian law and the law of armed conflict. With machines taking on the qualities of human intelligence, including perception, cognition and action, the issue of accountability for illegal acts performed by autonomous systems has dominated the debate. The issue is not so much *what* legal framework applies, but *who* is legally responsible for the effects of AI-enabled weapons.

AI-enabled action that constitutes an "armed attack" within the sense of Article 51 of the UN Charter, could potentially lead to the invocation of Article 5 of the North Atlantic Treaty. Setting aside the issue of the threshold to be applied to determine whether an AI-enabled attack would constitute an "armed attack", the preliminary question that would arise is that of attribution. After all, assigning responsibility for an armed attack is a necessary precondition to the use of self-defence measures. AI-enabled weapons challenge our traditional notions of responsibility: who can be held accountable for the effects of this technology, especially when fully autonomous? Can we even talk of a "legal personality" of AI-enabled systems? In this area, States have been rather clear in expressing the need for responsibility to be attributed to human beings or states actors.[15]

14 *Ibid.*, p.30.

15 See, for example, the reports of the Group of Governmental Experts 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". Reference is made to the *Report of*

For example, the international debates and discussions on the legal approach to take with respect to lethal autonomous weapons systems (LAWS) provides a useful reference on the concept of "human control", where the substance of the debate has been how to regulate human-machine interaction, often referred to as "human-machine teaming". The human role in independent machine decision-making can vary, as exemplified through the OODA loop of decision-making (Observe, Orient, Decide and Act).[16] An "in the loop" system requires human intervention for its operation, an "on the loop" system provides for human intervention if needed, and an "out of the loop" system does not require human intervention at all, a prospect that is criticized by observers who have supported the concept of "minimum human control". Privileging human judgment and accountability above mechanical efficiency, the United States, for example, has taken a clear stance towards asserting that a human being should always be kept in the decision-making loop for the use of LAWS, as the ultimate decider on the use of lethal force in the battlefield.[17]

Human intervention in the OODA loop may be a necessary condition for the application of international humanitarian law to the use of AI-enabled weapons. The rules of international humanitarian law apply to State actors who are behind the planning, the decision-making process and the execution of attacks. The four core principles of necessity, proportionality, discrimination, and humanity, are based on the assumption that human decision-making and judgment underlies military action: "these obligations cannot be delegated or given to machines, a computer program or a weapon system. It is individuals at the end of the day who are responsible for complying with International Humanitarian Law".[18]

the 2017 group of governmental experts on lethal autonomous weapons systems (LAWS), CCW/GGE.1/2017/CRP.1, 20 November 2017.

16 P. Tucker, "Tomorrow soldier: how the military is altering the limits of human performance", *Defense One,* 12 July 2017, https://www.defenseone.com/technology/2017/07/tomorrow-soldier-how-military-altering-limits-human-performance/139374/ (accessed 12 November 2019).

17 See United States Delegation Closing Statement to 2014 CCW Meeting of Experts on Lethal Autonomous Weapons Systems, 16 May 2014, (audio recording available at UN CCW 2014 LAWS meeting webpage) regarding the assertion that a human being should always be kept in the decision-making loop for the use of AWS, noting that the latter "does not sufficiently capture the full range of human activity that takes place in weapon system development, acquisition, fielding and use; including a particular commander's or an operator's judgment to employ a particular weapon to achieve a particular effect on a particular battlefield". The previous statement was found in: K. Neslage, "Does 'meaningful human control' have potential for the regulation of autonomous weapon systems", *National Security & Armed Conflict Law Review,* University of Miami, Vol.6, No.15, pp.151-177, https://nsac.law.miami.edu/wp-content/uploads/2015/11/Neslage-Final.pdf (accessed 12 November 2019).

18 Interview of the ICRC legal adviser Netta Goussac on the New technologies and IHL, ICRC Documents, 2018, https://www.icrc.org/en/document/icrc-legal-adviser-netta-goussac-talks-about-autonomous-weapons (accessed 12 November 2019).

Along the same lines, there is no particular reason to think that the traditional international law framework related to State responsibility could not be used in the context of military use of AI.

Conclusion

The aim of this chapter is to raise awareness and contribute to a shared understanding of the legal challenges raised by AI-enabled technology, particularly in the context of the international law applicable to collective defence. A primary assumption of this paper is that existing international law provides the appropriate legal framework, but that greater clarity is needed to determine some of the legal questions on how this existing legal framework applies to such new and evolving technologies, especially on the question of accountability. Going forward, States are encouraged to make their views known towards increased discussion and interaction. Such interaction would be useful for a variety of reasons, including to identify selected areas where national views and legal obligations amongst Allies may differ. Recognizing that States remain central to the development and the application of international law, NATO, as a multinational organization, can continue to provide a useful forum for State practice to be shared.

Conclusion

Catherine Warner

This volume has presented economic, psychological, historical, ethical, and legal insights into the use of Artificial Intelligence (AI) in society, and more importantly, in the defence domain. As pointed out in Chapter 5 by Marsan and Hill, "one of the core values that the Alliance defends is the rule of law". However, adversaries and resurgent threats are not restricted by such laws and are already using asymmetric warfare against us. NATO is currently facing a struggle with adversaries investing in new technologies, employing innovative tactics involving hybrid actions, deniable culpability, and anti-access area denial. This struggle is for economic control of new commercial routes opening up in the Arctic Ocean, as part of a quickly changing environment. It is a struggle prompted by the rush to exploit immense natural resources that were not thought of just a couple of decades ago. It is a struggle to control the info-sphere through cyberattacks. And in response, NATO forces must deny, contain, deter, and defend.

The Alliance's reaction to the current threat environment has to rely on innovation, on disruptive technological advances, and on renewed, stronger joint effort by Member Nations. But we must do this with the understanding that the current technological landscape is very different from the one the Alliance used to operate in. Today's emerging and disruptive technologies are driven by the commercial sector, characterized by an extremely short time to market; they are widely available and low cost. And these technologies can inflict damage not only to military forces but to civilian populations and infrastructure. NATO has to adapt to this new paradigm in order to maintain its technological edge.

As this volume looks specifically at the use of AI, one must "unpick" the meaning of this buzz-word and use a common lexicon to describe the risks and opportunities of this disruptive technology with a view to understanding its potential as a "game-changer." The concept of AI is not new. The field of AI research was founded at Dartmouth College, New Hampshire in the US, during the summer of 1956,[1] and has had a continually evolving definition over the decades as technology has matured. However, until access to large data

1 A. Kaplan and M. Haenlein, "Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence", *Business Horizons,* Vol.62, No.1, 2019, pp.15-25.

sets and the computing power to process them was available, the field did not progress much. Machine Learning, often depicted as a building block of AI, is the development of statistical models that make inferences based on large amounts of data. In the first decades of the 21st century, access to large amounts of data (known as "Big Data", or BD), faster computers, and advanced machine learning techniques were successfully applied to many problems throughout the economy and society. BD continues to transform our everyday lives and, as reiterated in this volume, 90 percent of data available in the world was created in the past 24 months. The availability of BD is what has spurred the development in the defence domain of image and video processing, text analysis, logistics management, and enabling processes such as automatic target classification or recognition.

The fields of unmanned and autonomous robotic systems are closely linked to AI, and they are often described together. They are perfect testbeds for demonstrating intelligent behaviour in physical systems and can be particularly useful for military applications. For example, it is generally agreed that unmanned systems can enhance situational awareness and reduce human workload, thus enhancing operational performance. The capabilities these systems bring to the fight are not necessarily superior to those provided by manned platforms, but they offer persistence, versatility, survivability, and reduced risk to human life. In the near future, unmanned systems are expected to become the operational planners' first choice for prolonged observation, such as airborne intelligence, surveillance and reconnaissance (ISR) or harbor security. Autonomous systems can execute missions that are dull, dirty, dangerous, or require prolonged endurance in harsh environments at a fraction of the cost of traditional assets, making them the perfect complement to traditional assets operating as a force multiplier.

Unmanned and autonomous systems will create fundamental shifts in the conduct of military operations. Consider the air domain, which has seen dramatic shifts in concepts of operations for air warfare with unmanned systems. These systems were not accepted immediately by air forces, as we read in Chapter 3, by Fino. One lesson other domains can learn from air is to not wait too long before establishing interoperability standards. As NATO nations continue to integrate unmanned systems into their armed forces, NATO must develop and test interoperability standards, thereby helping to de-risk the introduction of unmanned systems to the operational community and to reduce the techonological gap between Allies, as is mentioned in the Foreword.

Autonomous systems will not replace humans on the battlefield in the foreseeable future, but will complement and support them in their missions. This "system-of-systems" with humans must be experimented with and exercised as often as possible. In order to deliver science and technology solutions to the military, there must be close and continuing contact between the users and the scientific researchers. The requirements of the final customers (the operational community) play a crucial role in steering laboratory activities, making clear and open communication between the operational and scientific communities paramount. Wide-ranging conversations about the intended use of systems in development or the experiment being planned must flow between the customer and the lab, to ensure the continued relevance of scientific work.

Scientists must take new warfighting technologies to the operators, in their environment, to show the relevance of new technologies to their missions. I cannot emphasize enough how important this is. For the warfighters, and this new generation of operators, we want them to be comfortable and confident working with autonomous systems they can trust. We want them to know that scientists and engineers are building resilient systems that are trustworthy. We want sailors, soldiers, and airmen to trust autonomous systems so that they will use them.

However, the emergence of autonomous and intelligent technologies, particularly in a military context, raises important legal and ethical issues, as discussed in Chapters 4 and 5. Who is responsible for the decisions taken by a machine? At what point is the behaviour of a system reliable enough to be trusted for a particular mission? Can an autonomous system act in self-defence? These are examples of specific questions that must be considered. More broadly, we must assess the impact of new technologies on deterrence, escalation, or stability.[2] The decision to deny ourselves a technology that ensures our military superiority, while allowing potential adversaries to field it, must be evaluated from ethical, political, legal, and economic perspectives. The Alliance has been devoted to the settlement of international disputes by peaceful means since its creation. At the same time, Allies stand united in their commitment to deter any potential aggression and, if deterrence fails, to collectively defend themselves. Therefore, NATO must possess the full range of capabilities to fulfil its duty to deter and defend the citizens of the Alliance. Unfortunately, a real possibility exists that the Alliance could fall behind in defence Science & Technology in the coming years.[3]

2 T. S. Sechser *et al.*, “Emerging technologies and strategic stability in peacetime, crisis, and war”, *Journal of Strategic Studies*, Vol.42, No.6, 2019, pp.727-735; M. Horowitz, “When speed kills: lethal autonomous weapon systems, deterrence and stability”, *Journal of Strategic Studies*, Vol.42, No.6, 2019, pp.764-788.

3 L. Alleslev, *Maintaining The Edge & Enhancing Alliance agility*, NATO Parliamentary Assembly, Science and Technology Committee, Special Report, 2018, https://www.nato-pa.int/sites/default/files/2018-12/183%20STC%2018%20E%20-%20NATO%20SCIENCE%20AND%20TECHNOLOGY%20EDGE%20-%20ALLESLEV%20REPORT.pdf (accessed 12 November 2019).

A glimpse of what others are saying includes the following: in 2017, President Putin stated that "Artificial Intelligence is the future, not only for Russia but for all human kind. Whoever becomes the leader in this sphere will become the ruler of the world".[4] Meanwhile, China has launched a new AI megaproject. "Artificial Intelligence 2.0 will advance an ambitious, multibillion-dollar national agenda to achieve predominance in this critical technological domain, including through extensive funding for basic and applied research and development with commercial and military applications".[5] But, as pointed out in the introduction and in Chapter 1, technology cannot be a substitute for strategy, and future progress in AI is thus likely to reward those who first master its complements. NATO and the nations must think today how to interact with the machines of tomorrow.

4 K. Lant, "China, Russia and the US are in an Artificial Intelligence arms race", *Futurist*, 12 September 2017, https://futurism.com/china-russia-and-the-us-are-in-an-artificial-intelligence-arms-race (accessed 12 November 2019).

5 E. Kania, "Beyond CFIUS: the strategic challenge of China's rise in Artificial Intelligence", *Lawfareblog*, 20 June 2017, https://www.lawfareblog.com/beyond-cfius-strategic-challenge-chinas-rise-artificial-intelligence (accessed 12 November 2019).

NDC Publications (2018-2019)

NDC Policy Briefs

2018

The internal nature of the Alliance's cohesion
Thierry Tardy
NDC Policy Brief 1, October 2018

Projecting stability in practice? NATO's new training mission in Iraq
Kevin Koehler
NDC Policy Brief 2, October 2018

Challenges and potential for NATO-Egypt partnership
Adel El-Adawy
NDC Policy Brief 3, November 2018

The Great War legacy for NATO
Ian Hope
NDC Policy Brief 4, November 2018

European defence: what impact for NATO?
Thierry Tardy
NDC Policy Brief 5, December 2018

Will artificial intelligence challenge NATO interoperability?
Martin Dufour
NDC Policy Brief 6, December 2018

NATO-EU maritime cooperation: for what strategic effect?
Stefano Marcuzzi
NDC Policy Brief 7, December 2018

2019

Energy security in the Baltic Region: between markets and politics
Marc Ozawa
NDC Policy Brief 1, January 2019

NATO's nuclear deterrence: more important, yet more contested
Michael Rühle
NDC Policy Brief 2, January 2019

Vostok 2018: ten years of Russian strategic exercises and warfare preparation
Dave Johnson
NDC Policy Brief 3, February 2019 (8 pages)

Preparing for "NATO-mation": the Atlantic Alliance toward the age of artificial intelligence
Andrea Gilli
NDC Policy Brief 4, February 2019

NATO at 70: modernising for the future
Rose Gottemoeller
NDC Policy Brief 5, March 2019

NATO's coming existential challenge
Karl-Heinz Kamp
NDC Policy Brief 6, March 2019

"NATO@70": still adapting after all these years
Julian Lindley-French
NDC Policy Brief 7, March 2019

NATO is doing fine, but the Atlantic Alliance is in trouble
Bruno Tertrais
NDC Policy Brief 8, April 2019

70 years of NATO: the strength of the past, looking into the future
Kori Schake, with *Erica Pepe*
NDC Policy Brief 9, April 2019

NATO at 70: enter the technological age
Tomáš Valášek
NDC Policy Brief 10, April 2019

Building the airplane while flying: adapting NATO's force structure in an era of uncertainty
Sara Bjerg Moller
NDC Policy Brief 11, May 2019 (8 pages)

What NATO's counter-terrorism strategy?
Kris Quanten
NDC Policy Brief 12, May 2019

Why the Baltics matter. Defending NATO's North-Eastern border
Sven Sakkov
NDC Policy Brief 13, June 2019

The necessary adaptation of NATO's military instrument of power
Jan Broeks
NDC Policy Brief 14, June 2019

Deterring hybrid threats: the need for a more rational debate
Michael Rühle
NDC Policy Brief 15, July 2019

Russia and China: "axis of convenience" or a "stable strategic partnership"?
Marc Ozawa
NDC Policy Brief 16, July 2019

NATO and EU training missions in Iraq – an opportunity to enhance cooperation
Niels Schafranek
NDC Policy Brief 17, August 2019

Fighting "Men in Jeans" in the grey zone between peace and war
Peter Braun
NDC Policy Brief 18, August 2019

From hybrid warfare to "cybrid" campaigns: the new normal?
Antonio Missiroli
NDC Policy Brief 19, September 2019

The role of democracy and human rights adherence in NATO enlargement decisions
Eyal Rubinson
NDC Policy Brief 20, September 2019 (8 pages)

What NATO contribution to the security architecture of the Indo-Pacific?
Jean-Loup Samaan
NDC Policy Brief 21, October 2019

The enhanced Forward Presence: innovating NATO's deployment model for collective defence
Christian Leuprecht
NDC Policy Brief 22, October 2019 (8 pages)

NATO at 70: what defence policy and planning priorities?
Patrick Turner
NDC Policy Brief 23, October 2019

Calibrating the scope of NATO's mandate
Thierry Tardy
NDC Policy Brief 24, November 2019

NDC Research Papers

Projecting Stability: elixir or snake oil?
Edited by Ian Hope
NDC Research Paper 1, December 2018

NATO's Futures: the Atlantic Alliance between Power and Purpose
Sten Rynning
NDC Research Paper 2, March 2019

A Strategic Odyssey: Constancy of Purpose and Strategy-Making in NATO, 1949-2019
Diego A. Ruiz Palmer
NDC Research Paper 3, June 2019

Russia's Military Posture in the Arctic - Managing Hard Power in a "Low Tension" Environment
Mathieu Boulègue
NDC Research Paper 4, July 2019

NATO and the EU. The essential partners
Edited by Gustav Lindstrom and Thierry Tardy
NDC Research Paper 5, September 2019

NATO
OTAN
SINCE
DEPUIS 1949