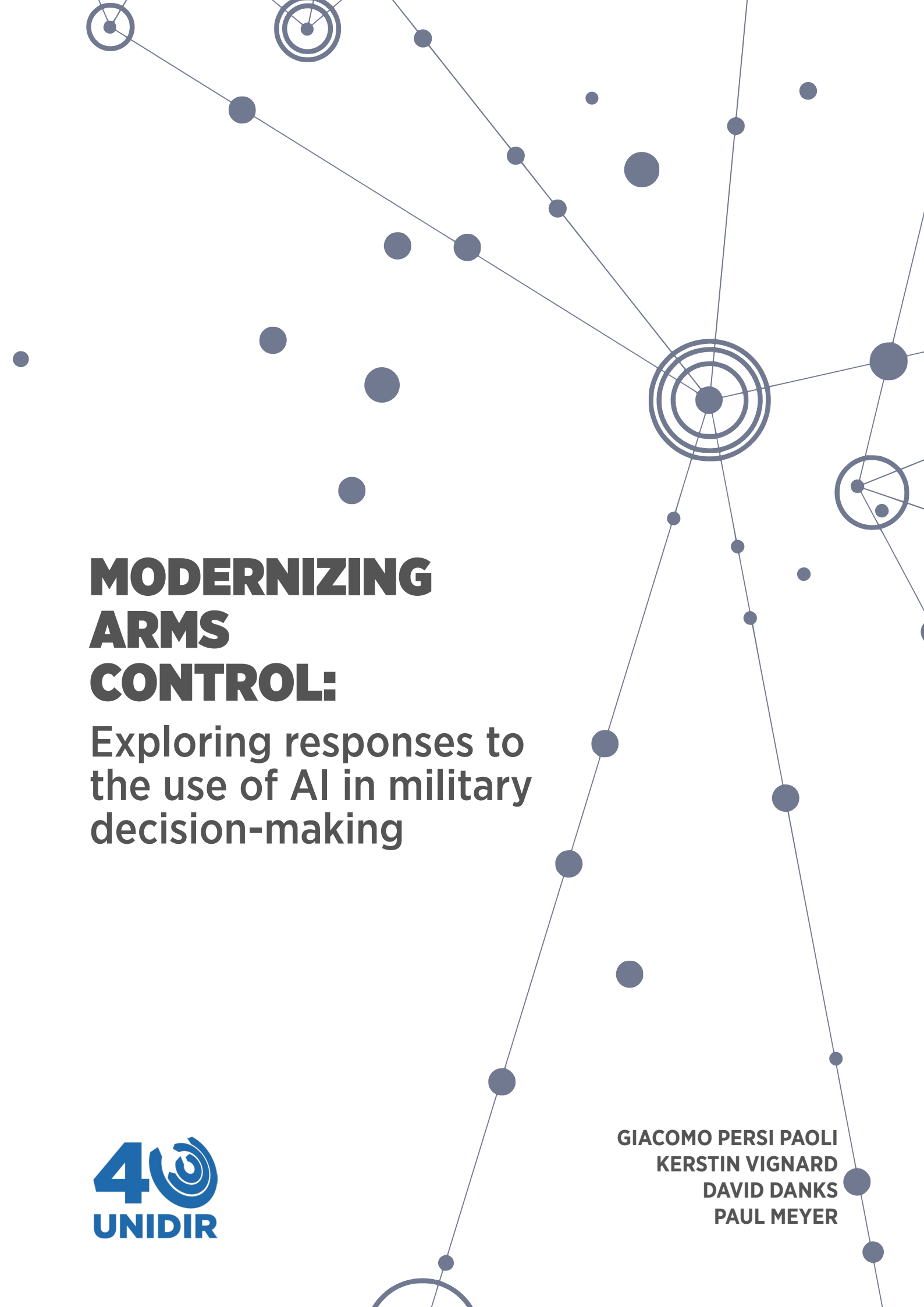


MODERNIZING ARMS CONTROL:

Exploring responses to the use of AI in military decision-making

GIACOMO PERSI PAOLI
KERSTIN VIGNARD
DAVID DANKS
PAUL MEYER

UNIDIR

ACKNOWLEDGEMENTS

UNIDIR and the authors would like to express their sincere gratitude to all the sponsors of this work.

UNIDIR would like to thank CIFAR, which through the Pan-Canadian AI Strategy, supported two expert workshops on the theme of this report (learn more about CIFAR at cifar.ca). UNIDIR also extends its appreciation to Osonde Osoba at the RAND Corporation for generously hosting the second workshop. The authors also wish to thank all the subject matter experts who participated in the workshops for their rich and invaluable input. The views expressed in this report are the sole responsibility of the individual authors.

This project was implemented within UNIDIR's Security and Technology Programme, which is funded by the Governments of Germany, the Netherlands, Norway and Switzerland, and by Microsoft. Support from UNIDIR core funders provides the foundation for all the Institute's activities.

Design and layout by Eric M. Schulz

NOTE

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries. The views expressed in the publication are the sole responsibility of the individual authors. They do not necessary reflect the views or opinions of the United Nations, UNIDIR, its staff members or sponsors.

CITATION

Persi Paoli. G., Vignard. K., Danks. D, and Meyer. P. "Modernizing Arms Control: Exploring responses to the use of AI in military decision-making". Geneva, Switzerland: UNIDIR.

ABOUT UNIDIR

The United Nations Institute for Disarmament Research (UNIDIR) is a voluntarily funded, autonomous institute within the United Nations. One of the few policy institutes worldwide focusing on disarmament, UNIDIR generates knowledge and promotes dialogue and action on disarmament and security. Based in Geneva, UNIDIR assists the international community to develop the practical, innovative ideas needed to find solutions to critical security problems.

www.unidir.org

TABLE OF CONTENTS

LIST OF BOXES

BOXES

ABBREVIATIONS AND ACRONYMS

AI	artificial intelligence
CBM	confidence-building measure
CCW	Convention on Certain Conventional Weapons
ICT	information and communications technology
ISO	International Organization for Standardization
OSCE	Organization for Security and Co-operation in Europe

ABOUT THE AUTHORS

Giacomo Persi Paoli is the Programme Lead for Security and Technology at UNIDIR. His recent work has focused on arms control, technology horizon scanning, artificial intelligence and cybersecurity. Before joining UNIDIR, Persi Paoli was Associate Director at RAND Europe, where he led the defence and security science, technology, and innovation portfolio as well as RAND's Centre for Futures and Foresight Studies. He served for 14 years as a Warfare Officer in the Italian Navy and has been extensively engaged in small arms and light weapons research in support of United Nations processes.

Kerstin Vignard currently leads the UNIDIR team supporting the Chairmen of the latest Group of Governmental Experts on Cyber Security and the Open-ended Working Group on information and communication technologies and international security. She has led UNIDIR's team supporting four previous cyber Groups of Governmental Experts. From 2013 to 2018, she established and led UNIDIR's work on the weaponization of increasingly autonomous technologies, which focused on advancing the multilateral discussions on weaponized autonomy by refining the areas of concern, identifying relevant linkages and learning from approaches from other domains of relevance (including the private sector). She is an international security policy professional with 25 years' experience at the United Nations and interests at the nexus of international security policy and technology. Vignard's areas of expertise include artificial intelligence, autonomous technologies, cyber, tech innovations and international security. Before her current special assignment, Vignard was Deputy to the Director at UNIDIR.

David Danks is the L.L. Thurstone Professor of Philosophy and Psychology and Head of the Department of Philosophy at Carnegie Mellon University. He is also an adjunct member of the Heinz College of Information Systems and Public Policy and the Center for the Neural Basis of Cognition. His research interests are at the intersection of philosophy, cognitive science and machine learning, using ideas, methods and frameworks from each to advance our understanding of complex, interdisciplinary problems. Danks has examined the ethical, psychological and policy issues around artificial intelligence and robotics in transportation, health care, privacy and security. He has also done significant research in computational cognitive science, culminating in his book *Unifying the Mind: Cognitive Representations as Graphical Models* (2014, MIT Press). Danks is the recipient of a James S. McDonnell Foundation Scholar Award, as well as an Andrew Carnegie Fellowship.

Paul Meyer is a Fellow in International Security and an Adjunct Professor of International Studies at Simon Fraser University in Vancouver. He is also a Senior Advisor with ICT4Peace and the Chair of the Canadian Pugwash Group. A former Canadian career diplomat, Meyer has served in a variety of positions abroad and in Ottawa with a focus on international security policy. He was Canada's Ambassador to the United Nations and the Conference on Disarmament in Geneva (2003–07). His research interests include international cyber security diplomacy, outer space security and nuclear arms control and disarmament.

EXECUTIVE SUMMARY

This report provides an initial insight into why the international security community may need to consider regulating artificial intelligence (AI) applications that fall in the digital grey zone between AI-enabled weapon systems (e.g. lethal autonomous weapon systems) and military uses of civilian AI applications (e.g. logistics, transport). It also provides an initial exploration of the familiar tools the community has at its disposal for such regulation.

Attempts to leverage arms control approaches to address digital technologies is not a new phenomenon. Since the first Russian-sponsored General Assembly resolution in 1998 on developments in the field of information and communications technologies and international security, states have grappled with how to adapt the concepts and tools of arms control to digital technologies – from the legality of offensive and defensive cyber operations, to non-proliferation measures aimed at encryption and particular types of software, to debates on whether to ban or regulate autonomous weapon systems.

Using the example of AI-enabled military decision support tools, this report offers an initial consideration of whether arms control approaches can enhance stability and reduce risks as military applications of AI become more widespread. Although it points to many areas where the "arms control toolbox" remains relevant or has potential to adapt, it also reaffirms the need for reconsideration of a much broader set of questions concerning how arms control is relevant to digital technologies, including definitional issues (What is an "arm"?), the utility of physical control measures, whether future efforts should address regulating objects or behaviour, and who are the relevant stakeholders necessary for effective responses.

UNDERSTANDING THE CHALLENGE

The risks that AI-enabled military decision support systems pose to international stability and security can be grouped into three categories: limitations in the technology itself, limitations in the humans tasked to use it and limitations in the environment in which it will be deployed.

- Limitations in the technology:
  - ✓ AI-enabled decision support systems often **lack clear (or clearly articulated) success criteria**.
  - ✓ Even when there are clear success criteria, algorithms **encode historical constraints and biases**, whether because they were provided with biased training data or used biased (or incorrect) assumptions, or for many other reasons.
- Limitations in human users:
  - ✓ **Humans frequently struggle** to use AI outputs correctly, whether owing to **automation bias** (where people rely on the AI output without questioning it) or **algorithmic aversion** (where people reject AI outputs simply because they came from an algorithm).
  - ✓ Significant research on the human–AI interface and the **education of users and operators** will be crucial for the relevant human decision makers to understand, trust or use AI outputs.
- Limitations in the environment:
  - ✓ AI-enabled military decision support systems will likely be deployed **in adversarial environments**. There are many demonstrations of adversarial "attacks" against AI systems that lead to significant misclassification or misprediction.
  - ✓ AI-enabled military decision support systems will likely be used in complex environments in which **unintended interactions** between systems may occur in unexpected or undesirable ways.

EXPLORING THE ROLE OF ARMS CONTROL FOR AI

Arms control measures do not exist in a vacuum but require an enabling environment in order to be effectively implemented, particularly if such measures are voluntary in nature and not legally binding. Strained relations between great powers, weakened multilateralism, erosion of existing regimes and agreements, and increased competition for political, military, and economic advantage are all factors that influence the impact or feasibility of certain tools and measures.

Among the vast literature and different views on the objectives of arms control, the following four arms control objectives are particularly salient:

1. **Stability:** To remove incentives for a first attack, to prevent accidental war and to reduce the risk of military escalation through enhanced predictability and transparency
2. **Safety:** To reduce the risks attendant upon military operations
3. **Legality:** To ensure compatibility with international legal obligations, notably international humanitarian law and human rights law
4. **Efficacy:** To provide sufficient incentives, and therefore good prospects, for the controls to be implemented and the desired conduct on the part of concerned states to be produced

While all four of these objectives are relevant when considering military applications of AI, the advent of AI in military decision-making blurs, in particular, the line between **safety** and **stability**: in traditional physical systems safety is associated with risk, while stability is linked to escalation, which is implicitly considered as intentional behaviour built on the assumption that humans are the ultimate decision makers. With increasing authority being delegated to AI systems, of which human operators may know too little, escalation may not be assumed to be "intentional" in the same way and may become the unintended result of unpredictable behaviour of algorithms.

THE ARMS CONTROL TOOLBOX

While in theory no arms control tool has to be excluded by default to achieve these objectives, in practice – given the dual-use nature of the technology, the range of actors involved and the current political environment – "softer" tools may present greater opportunities for success in the short term, despite some of their inherent limits (e.g. being potentially hard to verify and enforce). This is further reinforced by the increasingly critical role played by the private sector in the field of AI, in many cases replacing academia as the main repository of the non-governmental technical and scientific expertise needed to meaningfully advance arms control discussions.

Key insights emerging from the investigation of four traditional arms control tools can be summarized as follows:

» **International instruments, treaties and conventions**
  - In the last two decades, attempts to apply traditional arms control approaches to digital technologies have struggled. It is unlikely that a new, single multilateral instrument regulating the military use of AI will be negotiated in the near future.
  - However, existing instruments and commitments can be leveraged to introduce restrictions on military AI applications. For example, Article 36 of the 1997 Additional Protocol I to the Geneva Conventions may be particularly relevant. States could agree to incorporate AI-enabled decision support systems into the scope of "means or methods of warfare", subjecting them to the requirement of legal reviews outlined in this article.

» **International, regional and national export controls**
  - While it would be hard to enforce export controls on algorithms, a possible solution includes regulating AI chips, the technology required to manufacture them or the cloud infrastructure supporting them.

- Another option would be to focus on specific applications (e.g. audio and video manipulation technologies or decision support) rather than on their enabling technologies.

» **Other national controls**
- The development of a dedicated national directive could be used to create common baseline knowledge between national users, suppliers and developers of military applications of AI, including those supporting decision-making.
- This would include, for example, providing clear and shared definitions of concepts and assigning clear levels of responsibility to different actors throughout the life cycle of any given AI military application (e.g. from development, to testing, fielding and employment).

» **Voluntary measures: norms and standards**
- International, regional and national norms and standards are very powerful tools to promote the responsible development and safe adoption of a technology by states.
- In the field of AI, the most significant achievement so far, at least at the multilateral level, has been the endorsement by the High Contracting Parties to the Convention on Certain Conventional Weapons in 2019 of 11 guiding principles.
- An area in which the transferability of good practices from the cyber and information and communications technologies sector could be explored is the **development of standards**. Given the highly technical nature of the issue, **industry is likely to take a leading role** in the development of such standards, but it is important that states are involved in this process in order to maximize prospects for adoption.
- Finally, **codes of conduct** and **principles**. In the field of AI, many initiatives led by industry, national agencies or international organizations have approached "soft" regulation by means of AI principles. Leveraging these commitments and expressions of good intentions at the multilateral level remains an underused tool in international security discussions.

» **Confidence-building measures**
- Confidence-building measures (CBMs) are voluntary measures designed to prevent hostilities, to avert escalation, to reduce military tension and to build mutual trust between countries or communities.
- A relatively accessible CBM in support of transparency would be the development and public release of the national regulatory framework for AI-enabled military decision support (e.g. AI strategies, policies, directives and guidelines). This would also provide a baseline or blueprint for other governments in the process of developing their own regulatory frameworks.
- However, as different stakeholders play distinct roles within the AI ecosystem, it is important to differentiate between types of measure that could be designed for different types of interaction (e.g. continued engagement in government-to-government dialogue, joint development of standards between governments and industry, establishment of a neutral international scientific research centre [similar to CERN], or less institutionalized exchanges of knowledge to engage with the scientific community).

THE WAY FORWARD

A number of key takeaways can be extracted from this report, acknowledging that some of them might be relevant for multilateral discussions that go beyond the specific application of AI to military decision support and include wider applications of digital technologies (from cyber to autonomous weapon systems). In particular:

1. **AI in military decision support systems is an area that deserves further attention** from the international security community in order to manage risk and potential for instability
.

2. The traditional **objectives of arms control (stability, safety, legality, efficacy) remain valid and applicable** even when dealing with AI-enabled decision support systems.

3. The traditional arms control toolbox (e.g. international instruments, export control, voluntary measures, CBMs) will not become obsolete if the arms control community is open and willing to **embrace new forms of collaboration as well as adapt traditional ones** to fully **leverage the know-how of the scientific expert community, most of which now resides in the private sector**. This also entails creating more opportunities for exchanges and cross-fertilization of ideas and perspectives by involving industry and scientific experts in relevant arms control discussions as well as by ensuring that the arms control community actively engages with such actors in their relevant forums.

4. **There is no "one stop" solution**. A web of responses and incentive structures that target – and draw on the expertise of – different stakeholder groups will be required to effectively respond to the challenges posed by embedding AI in decision support applications.

5. The **range of possible measures** described in this report **does not always require government leadership** nor, in some cases, government direct participation (e.g. industry-led standardization processes, scientific knowledge exchanges). However, for these measures to produce meaningful impact on strategic stability and security, they would require recognition and downstream support by state actors. **Industry** can play a critical role provided it is given the opportunity to meaningfully engage with the arms control community.

6. Building on this last point, as thought-leader in AI, **industry has its own responsibilities**: from including legal and ethical considerations in their innovation policies and practices applied to AI, to be willing to consistently engage with regulatory processes at the international and national levels. This is particularly relevant for industry developing AI directly for defence, or for the broader industrial base developing potential dual-use AI algorithms or applications. Current efforts by a range of private sector actors to develop standards, or principles of responsible development of AI applications (e.g. transparency, reliability, security) are particularly relevant for the current debate on the international security implications of AI.

While the arms control community has been focused for the past several years on autonomous weapon systems, consideration of the international security dimension of AI-enabled decision support tools should also be part of the community's deliberations. A practical contribution to this endeavour would be additional research on the international security implications of technical aspects of algorithmic decision-making (e.g. explainability, predictability) in both weapon and decision support tools, as well as further exploration of softer regulatory approaches with greater involvement from the research and technical communities, as well as industry.

1. CONTEXT

Advances in artificial intelligence (AI) promise revolutionary benefits, optimization and competitive advantage in every sector, from health care and transportation, to entertainment and agriculture, to education and policy-making. Around the world, we are witnessing a period of massive public and private investment in AI as well as a growing number of national AI strategies and road maps indicating objectives and targets for the years to come.[1]

This surge to seize the promise of AI coincides with a period of growing international tensions and erosion of trust between major powers. Unsurprisingly, global competition to harness ever more powerful AI is also impacting existing great power rivalries, as reflected in changing national defence and security strategies, as well as the perception of who will be the key stakeholders in this increasingly strategic domain.

This growing competition is frequently described in the language of conflict and warfare. An increasingly common description is that the superpowers are engaged in an "AI arms race", as a consequence of the national interest both in the defence and security applications of AI and in its perceived strategic importance.[2]

In the strictest sense, an "arms race" refers to a competition between two or more states in the pursuit of military superiority.[3] It traditionally has both a quantitative dimension (number of armaments or forces) and a qualitative dimension (superior technology). The term "AI arms race" is widely used in a much looser sense to refer to fierce global strategic competition in AI. It may, for example, refer to the technological dominance of a company, the ability to leverage AI in increasingly sophisticated offensive and defensive cyber operations, the use of algorithms to manipulate human behaviour through news feeds and deep fakes, or the use of AI to enhance the capabilities of weapon systems. Simply put, AI is not a weapon, it is an enabling technology.[4]

1.1 MILITARY INTEREST IN AI

As an enabling technology, AI can be used to enhance or amplify existing capabilities and resources or to exploit data to identify patterns or make predictions or recommendations.[5] As such, AI systems can offer benefits to nearly every application: logistics, navigation, resource allocation, recruitment, health monitoring, translation, content analysis and more. In this way, military interest and investment in AI is no different than that of the civilian sector. Technological advances in autonomous vehicles, for example, could be deployed for personal transit, scientific exploration, humanitarian relief, medical evacuation or military convoys.

There is a wide spectrum of AI applications that militaries and defence forces are keen to leverage. On one end of the spectrum are those same applications as in the civilian sector, such as logistics, translation, image recognition, navigation or health diagnostics. At the other end of the spectrum are potential AI-enabled applications unique to the defence sector, such as physical weapon systems or offensive cyber operations, and the cyber–physical

1 In 2017 Canada launched development of the world's first national AI strategy (the Pan-Canadian Artificial intelligence Strategy). For an overview of national AI policies, see OECD AI Policy Observatory (2020).

2 While earlier uses of the term "AI arms race" have been documented, the trend of describing global AI competition as an arms race in the media started accelerating in 2015. For a sample of such reports in the popular press, see Apps (2019); Cohen (2017); Hughes (2017); Pecotic (2019); Scharre (2020); Thompson & Bremmer (2018).

3 While there is no universally accepted definition of an "arms race" or even "arms control", there is a well-developed literature on different definitional approaches and metrics. See, for example, Hammond (1993). For an introduction to the basic concepts of arms control, including its objectives, definitions and history in the modern age, see Goldblat (2002, 1–47).

4 AI is a "general purpose technology", commonly defined as "a new method of producing and inventing that is important enough to have a protracted aggregate impact" (Jovanovic & Rousseau, 2005). Other general purpose technologies include electricity and the steam engine.

5 For a general introduction to AI written for arms control policy makers, see UNIDIR (2018b).

systems that support them.[6] In the middle of the spectrum lie a range of AI-enabled applications applied to subjective or predictive tasks that support decision-making and analysis. These too have both civilian and military applications.

1.2 AI POLICY REGULATION

The AI policy regulation landscape is far from homogeneous, varying not only by country but also by application. Even within countries, local governments have taken very different approaches to different AI-enabled applications, such as autonomous vehicles and facial recognition. In many countries, debates about the regulation of AI-enabled applications have been characterized as pitting those who embrace "permissionless innovation"[7] against those favouring the precautionary principle.

Equally vast is the variety of regulatory responses being adopted, from hard measures, such as bans in some locations on particular uses of facial recognition, to soft measures, such as the development of IEEE P7000 standards by a global body of experts.

The most common narrow applications of AI, while dual use, are largely developed by the private sector for non-military applications. Even in the absence of specific legislation or regulatory frameworks, when concerns arise about these applications, they are scrutinized nationally and internationally by researchers, consumer groups and advocacy or rights organizations, and the producer or operator is under public pressure to respond. For example, demands for algorithmic fairness, accountability and transparency have pressured corporations and governmental bodies to modify, mitigate, improve or halt the use of particular AI-enabled applications. Such pressure has, for example, raised awareness of gender and racial bias in data sets and resulted in improvements in voice recognition and image classification. These improvements benefit both civilian and military applications.

AI-enabled weapon systems and AI-enhanced cyber operations are the subject of ongoing multilateral arms control discussions at the United Nations.[8] This report considers, instead, AI in military decision-making and support systems and whether the use of AI in such systems could raise novel risks and potential hazards to international security. Decision support applications are not weapon systems in themselves and thus have not yet been the subject of international arms control efforts. An area of early debate has been the potential benefits and risks of deploying AI in nuclear command and control systems.[9] Other areas include using predictive analytics in military detention or targeting decisions, similar to how algorithms are already deployed (and contested) in criminal justice and domestic policing applications in some countries.[10]

Attempts to leverage arms control approaches to address digital technologies is not a new phenomenon. Since the first Russian-sponsored General Assembly resolution in 1998 on developments in the field of information and communications technologies (ICTs) and international

6 Harmful or malicious uses of AI applications by non-state actors, including individuals, proxies or criminals groups, are beyond the scope of this report.

7 Thierer (2016) defines permissionless innovation as "the notion that experimentation with new technologies and business models should generally be permitted by default. Unless a compelling case can be made that a new invention will bring serious harm to society, innovation should be allowed to continue unabated and problems, if any develop, can be addressed later". Supporters of this concept believe that the precautionary principle disincentivizes innovation, lowers the quality of goods and services, and may negatively impact economic growth.

8 Lethal autonomous weapon systems, sometimes called "killer robots", have been the subject of expert discussion since 2014 within the framework of the Convention on Certain Conventional Weapons; see UNOG (2020b). The use of information and communications technologies in the context of international security (or "cyber warfare") has been the subject of negotiations within the United Nations framework since 2004; see UNODA (2020a).

9 Some military officials have announced clear "red lines" for the deployment of AI within nuclear command and control. Director of the US Department of Defense Joint Artificial Intelligence Center Lt. Gen. Jack Shanahan recently stated: "You will find no stronger proponent of integration of AI capabilities writ large into the Department of Defense … but there is one area where I pause, and it has to do with nuclear command and control," see Freedberg Jr. (2019). For an overview of considerations of AI in nuclear command and control, see for example Borrie (2019); Reiner and Wehsener (2019).

10 See for example Deeks (2018); Partnership on AI (2019).

security,[11] states have grappled with how to adapt the concepts and tools of arms control to digital technologies – from the legality of offensive and defensive cyber operations, to non-proliferation measures aimed at encryption and particular types of software, to debates on whether to ban or regulate autonomous weapon systems.

Using the example of AI-enabled military decision support tools, this report offers an initial consideration of whether arms control approaches can enhance stability and reduce risks as military applications of AI become more widespread. While it points to many areas where the "arms control toolbox" remains relevant or has potential to adapt, it also reaffirms the need for reconsideration of a much broader set of questions concerning how arms control is relevant to digital technologies, including definitional issues (What is an "arm"?), the utility of physical control measures, whether future efforts should address regulating objects or behaviour, and who are the relevant stakeholders necessary for effective responses.

This report considers whether particular military AI applications may create particular risks to international stability that we normally turn to arms control to address and identifies an initial set of potential responses. Chapter 2 describes why this set of military applications might raise novel (or compound existing) stability concerns not yet addressed through existing frameworks or controls. Chapters 3 and 4 consider the potential for arms control to help address or mitigate these risks. Chapter 5 provides a synthesis of the way forward.

11 UNGA (1999a).

2. FRAMING THE PROBLEM

2.1 EXPLORING THE BENEFITS

At present, AI systems are predominantly used for narrowly defined tasks in which there is a clear way of assessing success. For example, an image either does or does not contain a dog, the patient either does or does not have cancer, the AI either does or does not win the game of Go. AI is often employed to identify novel solutions in situations where humans do not know how to achieve success or the required rates of success. But present-day AI systems all depend on precise measures of performance, and many of the uses that have generated controversy – such as using AI to detect hate speech or predict recidivism risk – are problematic partly because there are not widely agreed-on or easily measured standards for assessing "success". For example, the use of risk assessment algorithms and machine learning in predictive policing efforts has revealed significant cultural biases within communities and has resulted in costly litigation and even lethal mistakes.

As noted above, AI systems are rapidly spreading across a range of military contexts and applications. Much of the attention on military AI systems has focused on weapons themselves, primarily on autonomous weapon systems. International arms control discussions on increasingly autonomous weapons have been under way at the United Nations since 2014, alongside multiple efforts by diverse groups to develop standards, guidelines or regulations governing or prohibiting such systems. However, what has been largely absent from the arms control discussion is attention directed to other AI-enabled military applications that support weapon systems and decisions to use them.

These applications include AI performing classification and prediction tasks, including signals or intelligence processing and decision support. In many contexts, these tasks are integral parts of processes that can lead to targeting or other decisions to exercise lethal force. These tasks all appear to have clear success criteria – the individual either is or is not a legitimate target, the vehicle either is or is not a tank, and so on – and so AI systems appear to be well suited to them. Decision support AI systems may provide guidance and insight to decision makers in, for example, command and control, including tactical and strategic planning; adversary prediction; targeting decisions; and information, surveillance and reconnaissance. For the purposes of this report, these applications are referred to as *military decision support AI systems*, or as *AI-enabled military decision support systems*.

AI-enabled decision support systems have the potential to provide enormous benefits to militaries. For example, modern militaries often have access to large-scale information, surveillance and reconnaissance sensor data that cannot be efficiently or effectively processed by humans alone and require significant AI assistance. Alternately, targeting decisions could be readily supported by image or video classification and tracking AI systems. While recognizing the potential benefits of deploying AI systems for these tasks, it is necessary to also consider if such applications may present risks and, if so, what could be done to address those risks.

Advanced militaries around the world have publicly declared their intention to use AI systems to support intelligence analysis and planning and to more generally assist human decision makers in having better situational awareness and understanding. By way of example, the US Department of Defense recently requested proposals for AI elements in its Joint Warfighting National Mission Initiative,[12] partly to improve its soldiers' ability to understand and use sensor intelligence. Or consider the prominent role for AI in situational awareness and information superiority in the 2019 defence white paper that guides Chinese

12 JAIC (2019).

military development and acquisition[13] as well as the relevance of the establishment in India of the High Level Defence AI Council, which will provide strategic guidance toward AI-driven transformation in defence.[14] Precisely because these systems are not weapons and because they are frequently integrated into existing processes and workflows, one may expect to see them developed and deployed quite quickly and outside many of the normal, weapon-centric military acquisition processes. Because decision support AI systems are rarely understood as potentially problematic to the same degree as weapon systems, they are not typically subjected to the same kinds of oversight and testing as weapons.

One key reason why AI-enabled decision support systems have received relatively less attention from the arms control community is that these systems are always deployed in conjunction with a human decision maker. As such, they appear to present fewer of the legal or ethical concerns that have been raised about autonomous systems. However, AI-enabled decision support systems can lead human decision makers to choices that undermine stability and safety, potentially without corresponding gains in efficacy. For example, if an information, surveillance and reconnaissance AI system makes significantly distorted or incorrect classifications, then the human decision maker who receives that "information" could unknowingly make poor decisions. Of course, bad information has always negatively impacted decisions. But when significant aspects of the decision process are "offloaded" to an AI system, or when the human decision maker does not understand the capabilities and limitations of that system, there is much greater potential for unintentionally destabilizing decisions.

Even if these systems have received little attention from the arms control community, they have been a catalyst for tech industry workers and the research community. Their concern received international attention in 2018 when over 4,000 Google employees protested the company's work on a US Department of Defense project using AI to interpret footage from unmanned aerial vehicles, stating, "We believe that Google should not be in the business of war".[15] Other large tech companies, such as Amazon and Microsoft, have seen their employees question the ethics of working on defence contracts, including for facial recognition technology and cloud computing services. Many claim that they would never choose to work on a weapon system or for the military and believe that these sorts of application may ultimately support weapon targeting with decreasing human oversight or intervention.[16]

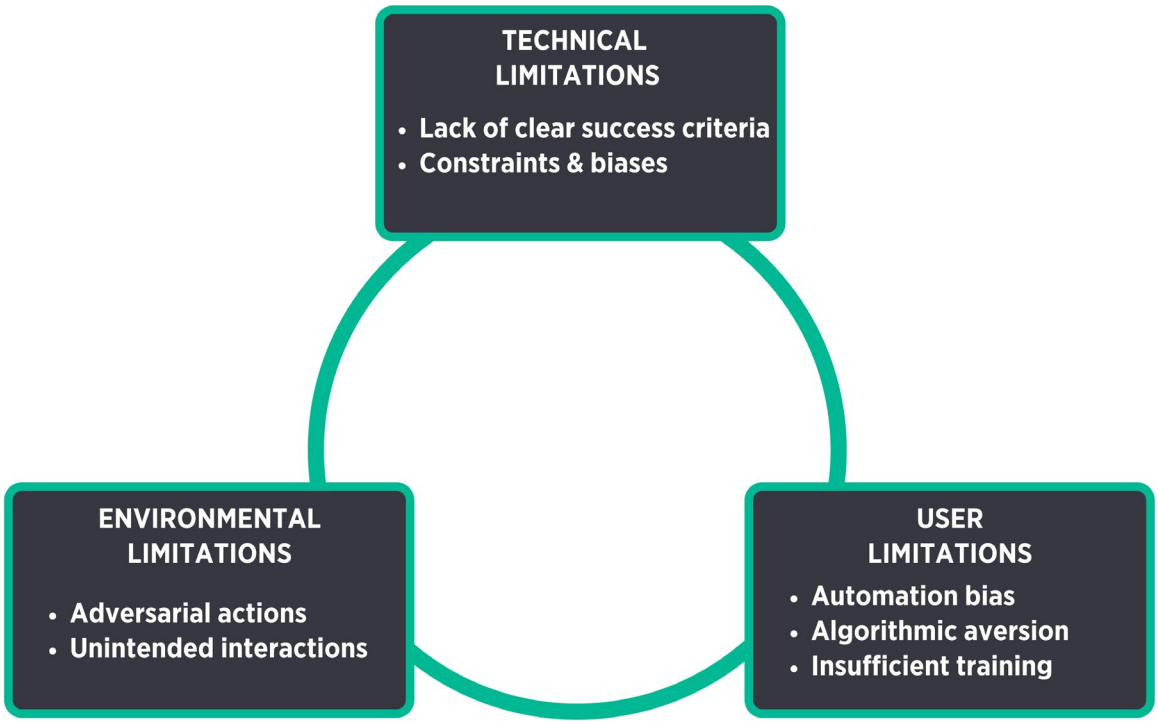
2.2 UNDERSTANDING THE RISKS

Recent years have revealed a wide range of limitations on using AI systems for classification and prediction tasks. Seven concerns in particular should be highlighted when considering the risks of AI-enabled military decision support systems. These concerns can be grouped into three categories: limitations in the technology itself, limitations in the humans tasked to use it and limitations in the environment in which it will be deployed (figure 1).

Many natural contexts for AI-enabled decision support systems actually **do not have clear (or clearly articulated) success criteria**. As a result, it is challenging for these AI systems to perform particularly well in specific tasks, even if all else works correctly. For example, consider the classification task of identifying individuals engaged in hostile activities. The category of "hostile action" does not have sharp boundaries, nor are there clear signals that always (and only) indicate such actions. The

13 State Council Information Office (2019).
14 Sarangi (2019).
15 See Shane and Wakabayashi (2018); Tiku (2018).
16 Tech worker protests have not been limited to military applications; domestic applications, particularly by law enforcement and border control agencies, and in particular the use of surveillance and facial recognition technologies, are the subject of increasingly vocal concern. Recent events, including high-profile incidents of racial injustice and the use of force in response to peaceful protests, have also served as a catalyst for broader societal conversations about the legality and ethics of particular uses of AI-enabled technologies.

FIGURE 1 Overview of Limitations Associated with AI-Enabled Decision Support Systems



same behaviour could be perceived by one person as hostile and by another as helpful, perhaps because they have different knowledge about social norms or customs. However, the AI system simply classifies and so will inevitably favour one or the other of these interpretations. In addition, AI systems in these contexts are more likely to encounter difficult edge cases or significant ambiguities, and these inputs are typically the most challenging for AI system performance. Unless there is substantial education of the human decision makers about limitations and imposed interpretations, decision support AI systems may contribute to a skewed understanding or awareness of the situation.

Even when there are clear success criteria, algorithms are now **recognized to encode historical constraints and biases**, whether because they were provided with biased training data or used biased (or incorrect) assumptions, or for many other reasons.[17] In some contexts, this connection to historical factors is the key to success, but in others, it can lead to quite problematic algorithmic outputs. For example, a predictive policing algorithm that is trained on data reflecting historic racially biased policing practices will replicate those biases in its subsequent predictions.[18] Algorithmic adjustments could compensate for some of these biases, but often only at the cost of introducing or exacerbating other kinds of bias. A designer, for example, might be able to reduce moral biases, but only by increasing statistical biases. Algorithmic biases and trade-offs in AI-enabled military decision support systems are similarly inevitable. Relatedly, algorithms can exhibit surprising and problematic errors if their use contexts are changed from the historical ones. In dynamic, fluid situations such as conflicts or disputes, the historical data and the resulting algorithm might not track the changing circumstances. These concerns are magnified given the complexities

17 The list of papers on algorithmic bias is much too long to include here. For an overview of algorithmic bias, see UNIDIR (2018a). Other early representative papers include Barocas and Selbst (2016); Danks and London (2017).
18 Ferguson (2017).

of data collection in adversarial and rapidly changing international environments; training data may be biased, unrepresentative or otherwise fail to capture key elements of the specific use context.

A second category of concern refers to limitations in the users. **Humans frequently struggle** to use AI outputs correctly, whether owing to **automation bias** (where people rely on the output without questioning it) or **algorithmic aversion** (where people reject outputs simply because they came from an algorithm).[19] In addition to the development of the system itself, significant research on the human–AI interface and the **education of users and operators** will be crucial for human decision makers to understand, trust and effectively use AI outputs.

A last set of concerns relates to the environment in which systems might be used. AI-enabled military decision support systems will usually be used to support operations **in adversarial environments**. There are many demonstrations of adversarial "attacks" against AI systems that lead to significant misclassification or misprediction. In two of the best-known examples, the application of apparent noise to an image led to a plastic turtle being classified as a rifle,[20] while the application of pieces of tape led to a stop sign being perceived as a yield sign.[21] In these and many other cases, adversarial attacks can produce significant errors in AI outputs. Moreover, these attacks can often be conducted in ways that are largely undetectable by humans. In a military context, for example, a decision support system providing intelligence about an area under an adversary's control could be tricked into producing quite incorrect classifications or predictions, thereby leading to significant potential harm to civilians or an incident of "friendly fire".[22]

Some of the environmental concerns are also that AI-enabled military decision support systems will likely be used in complex environments in which **unintended or undesired interactions** between systems can be expected. A compelling example of the risks associated with unintended interaction between algorithms is provided by

19 For example, see Albright (2019); Dietvorst et al. (2015); Skitka et al. (2000).
20 Athalye et al. (2018).
21 Eykholt et al. (2018).
22 Danks (2020).

algorithmic trading in the financial sector, which in 2010 played a crucial role in what became better known as the "Flash Crash". The official report by the US Securities and Exchange Commission and the Commodity Futures Trading Commission described how high-frequency trading algorithms automatically started executing buying and reselling orders, some at irrational prices as low as one penny or as high as $100,000.[23] As a result, investors witnessed nearly $1 trillion of value being erased from US stocks in a matter of minutes.[24] In the military context, these unintended interactions could arise within a single military's AI systems, between the AI systems of allies, or even between the AI systems of adversaries. The increased speed of AI systems compounds this risk, since unintended interactions may occur too fast for human decision makers to intervene.

Present-day AI systems, particularly those based on "deep learning", are notoriously difficult to explain or understand, despite "explainable AI" being a major research area.[25] The structure and training of these algorithms render them largely opaque to humans, whether users or developers. As a result, AI systems are often described as "black boxes": they are capable of very high-accuracy predictions and classifications, but from the outside their workings are completely mysterious. Methods have been developed to incrementally improve understanding and testing of AI systems, but those methods are quite intrusive and require a level of access to the systems themselves that is in significant tension with military secrecy and compartmentalization. Militaries usually tip the balance of transparency and secrecy toward the latter, and so AI systems could be considered "double black boxes": the black box of the AI system inside the black box of organizations that are rarely transparent. Mechanisms such as third-party oversight are currently being used for other governmental applications of AI, such as child welfare or judicial decision-making, but are unlikely to be feasible control or oversight mechanisms for many military AI systems.

There are significant ongoing efforts in civilian contexts to address some of the seven challenges articulated above. AI safety is an active field of research that seeks to address or mitigate these unintended or harmful behaviours. Non-military uses of AI in industry and government are increasingly subjected to significant formal and informal oversight, often leading to positive changes. Numerous corporations, particularly technology companies, are adopting principles or normative frameworks for ethical and responsible use of AI. Professional organizations such as IEEE and the Association for Computing Machinery are developing codes of conduct and professional ethics to guide AI developers. Section 4.3 of this report considers if these and other sorts of responses might be appropriate for and adapted to the context of military decision support AI systems.

23 US Commodity Futures Trading Commission and US Securities and Exchange Commission (2010).
24 Levine (2015).
25 Guidotti et al. (2019).


Ai

3. CONCEPTUALIZING ARMS CONTROL FOR AI

Originally, "arms control" was meant to denote rules of limiting arms competition (mainly nuclear) rather than reversing it. This term had a connotation distinct from "regulation of armaments" or "disarmament", the terms used in the United Nations Charter. Subsequently, however, a wide range of measures have come to be included under the rubric of arms control, in particular, those intended to: (a) freeze, limit, reduce or abolish certain categories of weapons; (b) ban the testing of certain weapons; (c) prevent certain military activities; (d) regulate the deployment of armed forces; (e) proscribe transfers of some militarily important items; (f) reduce the risk of accidental war; (g) constrain or prohibit the use of certain weapons or methods of war; and (h) build up confidence among states through greater openness in military matters.

Jozef Goldblat, Arms Control: The New Guide to Negotiations and Agreements

3.1 THE OBJECTIVES OF ARMS CONTROL

Among the vast literature and different views on the objectives of arms control, the following four arms control objectives are particularly salient:

1. **Stability:** To remove incentives for a first attack, to prevent accidental war and to reduce the risk of intended or unintended military escalation through enhanced predictability and transparency.
2. **Safety:** To reduce the risks attendant upon military operations.
3. **Legality:** To ensure compatibility with international legal obligations, notably international humanitarian law and human rights law.
4. **Efficacy:** To provide sufficient incentives, and therefore good prospects, for the controls to be implemented and the desired conduct on the part of concerned states to be produced.

Traditionally, arms control has addressed both hardware (i.e. reduction or elimination of actual weapons) and behaviour (i.e. constraints on the deployment of military forces or equipment), but emphasis has been on agreements dealing with specific weapon systems. In part, this reflected the need to ensure verification (see box 2) of the agreements, which was more readily accomplished when a physical object could be monitored with a high degree of accuracy.

However, these four objectives are desirable and remain relevant in the context of military decision support systems that are not weapons in themselves.

There are several ways in which the increased use of AI in military decision-making could undermine **stability** by triggering escalation. For example, increased use of AI from certain military powers could encourage others to pursue weapons build-up and investment or first mover advantage in both the development and deployment of new systems and new capabilities (e.g. loitering munitions with increasingly long periods of autonomy, or increasingly autonomous underwater objects, that lower the risk and cost for the deployer but whose deployment could be perceived as aggressive). In addition, stability could also be undermined by unpredictable interactions between AI systems that lead to unintended escalation.

As a consequence, the objective of **safety and risk reduction** deserves particular attention as it relates directly to the field of AI safety described in the previous chapter. In simple terms, AI safety refers to the technical, regulatory and ethical principles designed to reduce the risk of accidents in machine learning systems, where accidents are defined as "unintended and harmful behaviour that may emerge from poor design of real-world AI systems".[26] Such accidents may be related to the

26 Amodei et al. (2016, 21).

performance of an individual AI system or to the interaction between different AI systems.[27]

Transferring this example of unintended behaviour to the context of military operations where AI systems may be used to increase the responsiveness of weapon systems or prioritize targets clearly highlights the importance of safe operations and the avoidance of unrestrained escalation. These risks are not restricted to lethal autonomous weapon systems but apply to decision support AI systems as well.

In addition, the advent of AI in military decision-making **blurs the line between safety and stability**: in traditional physical systems, safety is associated with risk, while stability is linked to escalation, which is normally implicitly considered as intentional behaviour built on the assumption that humans are the ultimate decision makers. With increasing authority being delegated to AI systems, of which human operators may know too little, escalation may not be assumed to be intentional in the same way.

From a **legality** perspective, the key question focuses on the non-transferability of the principle of accountability. Just as much of the international discussion on autonomous weapon systems has focused on debates over human control and responsibility, similar questions about responsibility may arise when human operators or users rely on or base their actions on the AI system analysis or recommendation.[28]

Finally, from an **efficacy** point of view, a key issue to consider is the dual-use nature of AI and its associated elements, from code to AI chips. Ensuring good prospects for the implementation of control measures, the achievement of desired outcomes, and the minimization of unintended negative externalities requires a system-level approach that takes into account incentive structures for all actors involved, including states, industry and the community of AI researchers and developers.

3.2 THE ACTORS OF ARMS CONTROL

Arms control is often understood as primarily a state-led system characterized by state-to-state negotiations of bilateral and multilateral agreements and arrangements through which states mutually agree to either limit the proliferation and regulate the use of weapons or prohibit them entirely.

While states are undeniably at the heart of arms control, history is replete with examples of the instrumental role that non-governmental stakeholders – non-governmental experts and professionals; civil society actors, such as advocacy organizations; industry groups; and others – have played at the domestic, national and transnational levels in advancing and realizing arms control objectives.[29]

Scientists, engineers and technical experts have long been influential actors in many arms control processes and negotiations.[30] The knowledge base that has been relevant to weapon development is also essential for devising the best means of controlling or eliminating these weapons and for, eventually, verifying these processes. Thus, states have called on their expertise through both formal and informal mechanisms during the negotiation (and implementation) of instruments relating to nuclear, biological and chemical weapons, as well as domain-specific instruments, such as those focused on the seabed or outer space.

27 For an overview of risk and safety concerns of algorithms embedded in military systems, see UNIDIR (2016).
28 UNOG (2020b).
29 Johnson (2011).
30 While this paper focuses on arms control for strategic stability, these non-state actors have played a critical role in the development and implementation of a range of instruments that put the humanitarian element at their core. For example, through rigorous political lobbying, awareness raising, collective activism, and stimulation of public conscience and public discourse, the International Campaign to Ban Landmines, global citizens and a variety of non-governmental organizations were instrumental in the adoption and signature of the Mine Ban Treaty. See Atwood (1999).

Scientific advisers have had formal roles in the development of all arms control instruments related to weapons of mass destruction. A few examples:

» The conclusion of the Comprehensive Nuclear-Test-Ban Treaty in 1996 was preceded by the work of the Group of Scientific Experts, established in 1976 to examine the optimal technologies needed to confidently verify an eventual nuclear test ban.
» A scientific and technical experts' group has also been a feature of the 1972 Biological and Toxin Weapons Convention from its inception. The group led work on verification options for the treaty, and scientists, academics and industry experts provided direct scientific and technical expertise to states parties vital for the implementation of the Convention.[31]
» The Organisation for the Prohibition of Chemical Weapons, the implementing entity for the 1993 Chemical Weapons Convention, employs hundreds of technical experts engaged in the research and verification work that supports the Convention.

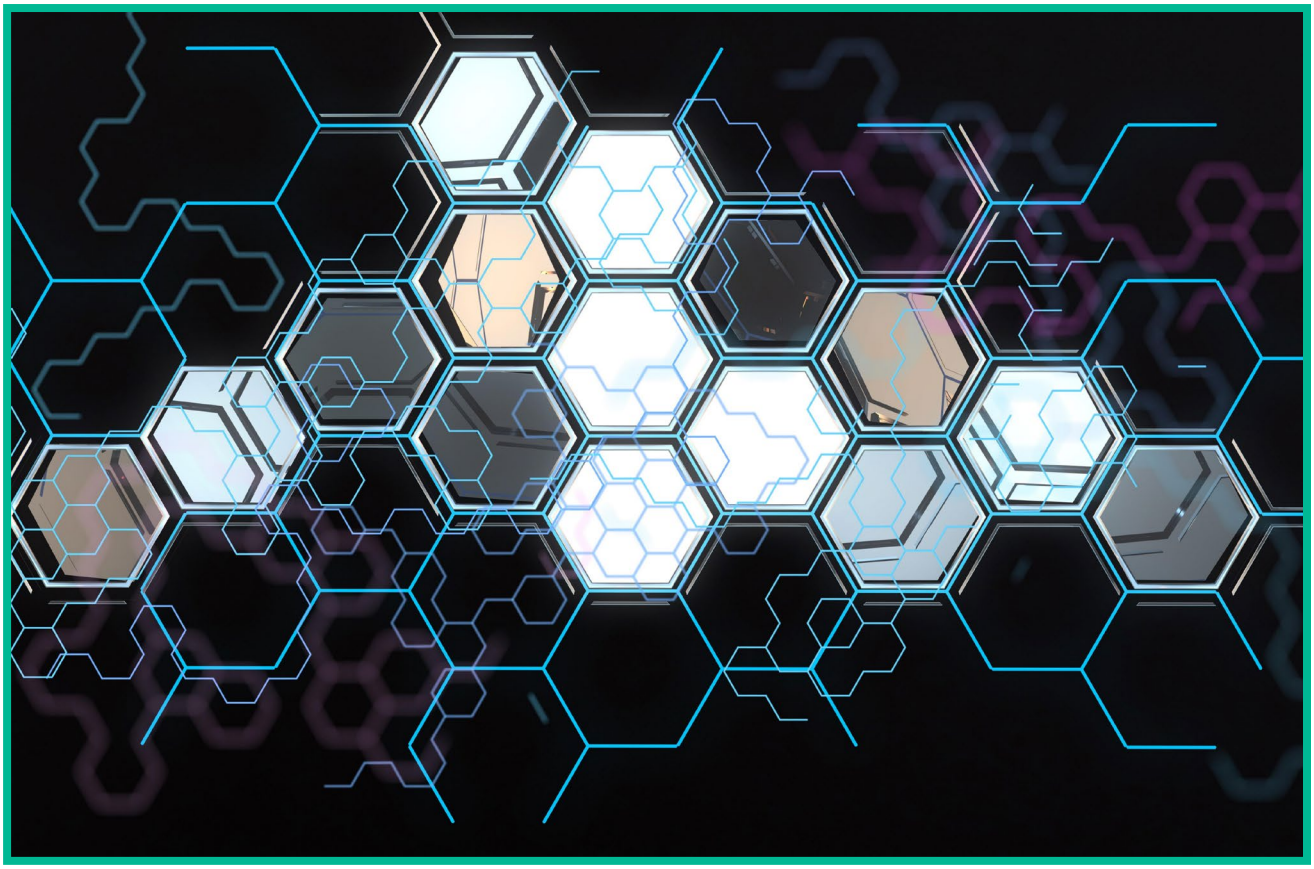
The link between treaty, science and technology that is a feature of so many arms control agreements is well expressed on the Organisation for the Prohibition of Chemical Weapons website: "The Chemical Weapons Convention (…) is built on a scientific foundation. Effective implementation requires technical expertise and scientific literacy for decision making."[32]

There is also a long history of scientific experts creating their own networks and organizations in support of arms control. The International Panel on Fissile Materials – composed of lawyers, scientists, nuclear experts and negotiation specialists – provides key research and expertise, including drafting options, for a fissile material cut-off treaty to support progress toward achieving a legally binding instrument.[33] In addition, many disarmament advocacy organizations have been founded by scientists and support scientific-evidence-based policy-making and "Track 1.5" confidence-building measures (CBMs; see section 4.4),[34] such as the Pugwash Conferences on Science and World Affairs, an organization whose origins are in the anti-nuclear Russell–Einstein Manifesto; the Bulletin of the Atomic Scientists, founded by Manhattan Project scientists; and International Physicians for the Prevention of Nuclear War. National science academies have also been active participants and supporters of various arms control regimes, and often their international projects have been Track 1.5 CBMs in their own right.

In the context of multilateral processes related to digital technologies and international security, somewhat counter-intuitively, the role of scientific experts and technologists has been much slower and more limited.[35] In the discussions on autonomous weapon systems within the framework of the Convention on Certain Conventional Weapons (CCW), some states have invited technical experts, mostly from national research establishments or universities, to serve as advisers on their delegations. Thus far, the CCW has no equivalent of, for example, the Group of Scientific Experts, which helped states understand the scientific techniques that could be best leveraged to verify a ban on nuclear testing. With no formal mechanism for engaging scientific experts in the multilateral discussions related to autonomous weapon systems, most scientific experts who have taken an interest in this field do so via one of the advocacy groups actively engaging with these discussions.[36]

Multilateral discussions on ICTs and international security have been even more closed to external expertise. It was only in

31 King's College London and Geneva Disarmament Platform (2017).
32 Pontes (n.d.).
33 Johnson (2011).
34 Refers to diplomatic tracks: Track 1 between government officials; Track 2 between non-governmental experts; and Track 1.5, a combination of government officials and non-governmental experts.
35 For an overview of why additional engagement with the technical community is necessary in arms control processes focusing on emerging technologies, see Vignard (2018).
36 Such as the International Committee for Robot Arms Control (see ICRAC [2020]) or the Campaign to Stop Killer Robots (see Campaign to Stop Killer Robots [2020]).



December 2019 that the first informal consultations were held between states and non-governmental stakeholders, including the private sector,[37] on the international security dimension of ICTs – an important yet overdue step in recognizing the critical role that non-government knowledge and expertise can play in advancing discussions and negotiations in the sophisticated and fast-paced digital sector.

Overall, the private sector's interest in, and engagement with, arms control processes on digital technologies has been slow. As an example, in the information technology sector a small number of major industry actors have begun to actively contribute to discussions at the multilateral level as well as champion the implementation of norms of responsible behaviour in cyberspace through mulitstakeholder and sector-specific initiatives.[38] In contrast, industry leaders in AI, or industry groups representing this sector, remain noticeably absent from discussions on autonomous weapons, perhaps out of concern about negative public perception.

This points to an important change from the historical engagement of scientific expertise in arms control discussions: whereas historically much of the relevant non-governmental scientific expertise was in academia, in AI a large part is in the private sector. Many states may feel that inviting technical specialists from the private sector to advise their delegation – or nominating them to serve as members of an international advisory committee – would be either a real or a perceived conflict of interest. Another complication in bringing industry into multilateral discussions is the fact that the relationship between states and industry differs across countries, by sector, and, sometimes, even at the individual company (e.g. state-owned enterprises, partial state participation, or fully private). This is an additional sensitive factor in considerations of how to engage industry in multi-stakeholder approaches at multilateral levels.

As such, 'engaging industry' is not as straightforward as it might otherwise be. However, if this remains the case, new modalities will need to be developed to bring this expertise to bear on discussions on international security and AI.

37 See, for example, Permanent Mission of Switzerland to the United Nations (2020).
38 See, for example, the Cybersecurity Tech Accord (2020).

4. THE ARMS CONTROL TOOLBOX AND AI: STRENGTHS AND LIMITATIONS

Traditionally, arms control has contributed to maintaining the "strategic stability" among adversarial states through a series of different measures, ranging from legally binding agreements reducing or constraining forces, to softer arrangements designed to build confidence by enhancing transparency and predictability. These two approaches address different aspects of the overall problem: one concentrates on the reduction or elimination of actual weapon systems; the other seeks to condition their deployment (and the deployment of armed forces).

BOX 1 Multi-stakeholder Approaches to International Security and Emerging Technologies

The international arms control community has acknowledged that addressing emerging technologies and international security requires a complementary and overlapping set of responses wielded by a range of actors. The most explicit articulation of this so far has been in the 2018 discussions on lethal autonomous weapon systems, where delegations recognized six stages in the context of emerging technologies in the area of lethal autonomous weapon systems: [1]

- Political direction in the pre-development phase
- Research and development
- Testing, evaluation and certification
- Deployment, training, command and control
- •Use and abort
- •Post-use assessment

National regulation, industry standards and international regulation were identified as three overlapping response mechanisms. Industry standards were considered more relevant to research and development and to testing, evaluation and certification, while international regulation was considered likely to come into play in deployment, training, command and control as well as use and abort. National regulation was recognized to have a role in all six stages.

Recognizing the importance of each response mechanism is a necessary step. However, much more needs to be done to strengthen links and coherence between them. How might interplay between them be fostered? For example, as technical bodies develop standards, how could awareness of international security concerns about dual-use applications be raised? Or how might the expertise of the AI safety research community be called on in arms control discussions? Without active promotion of such exchanges, including through establishing formal mechanisms for the regular participation of technical experts in arms control discussions and processes, such interactions are unlikely to happen on anything other than an ad hoc basis.

1 See the "sunrise slide", GGE (2018, 14).

How relevant are the traditional tools of arms control in the context of military applications of AI? There is no single answer to such a question, as in principle most traditional tools remain valid but require a different framing or approach to their implementation, as well as engagement with different partners and stakeholders.

Arms control measures do not exist in a vacuum and require an enabling environment to be effectively implemented, particularly if such measures are voluntary in nature and not legally binding. Strained relations between great powers, weakened multilateralism, erosion of existing regimes and agreements, and increased competition for political, military and economic advantage are all factors that influence the impact or feasibility of certain tools and measures. Therefore, the following sections describe different types of tool and, for each, provide an illustrative example of how such tools could be relevant to addressing the use of AI-enabled decision-making systems.

While in theory no arms control tool has been excluded by default, in practice – given the dual-use nature of the technology, its intangibility, the range of actors, and the current political environment – "softer" tools (see sections 4.3 and 4.4) may present greater opportunities for success in the shorter term, despite some of their inherent limitations (e.g. being potentially hard to verify and enforce).

4.1 INSTRUMENTS, TREATIES AND CONVENTIONS

Since the inception of arms control, the development of legally binding instruments has been, to a certain extent, technology driven, either in response to how technological advancements could be leveraged in the military domain and affect strategic stability and security or in response to new possibilities offered by technology in the context of compliance, trust and confidence (e.g. verification).

Arms control instruments can take the form of bilateral accords, as well as multilateral ones, with the objective of agreeing on limitations or measures of restraint in support of non-proliferation and disarmament. [39]

Initial attempts to apply traditional arms control approaches to digital technologies have struggled. For example, United Nations Member States have been discussing the international security dimension of ICTs for over two decades, with slow and limited success.[40] Attempts to regulate or control computer code or algorithms as potential "cyber weapons" have not resulted in a single legally binding arrangement or even a definition of the items or tools of concern (i.e. what constitutes a cyber weapon). Another key challenge related to digital technologies is reflected in the difficulty of designing an effective mechanism for monitoring compliance and enforcement of any hypothetical treaty or convention. With many states opposing attempts to develop a binding instrument, focus has been on affirming that existing international law applies to state use of ICTs, agreeing to 11 voluntary norms of responsible behaviour in the use of ICTs by states, and promoting a variety of confidence-building and cooperative mea-

39 For example, an initial arms control measure that limited the testing of nuclear weapons to underground sites, the Limited Test Ban Treaty, was concluded in 1963. Bilateral agreements of ever-increasing scope followed, with the Anti-Ballistic Missile Treaty of 1972 and an alphabet soup series of SALT, START, SORT and New START treaties limiting the levels of strategic weaponry. In addition to the bilateral accords, multilateral agreements emerged with the conclusion of the Outer Space Treaty of 1967 (prohibiting weapons of mass destruction in space) and the 1968 Nuclear Non-Proliferation Treaty, enshrining both non-proliferation and disarmament obligations.

40 From 2004 until 2017, the United Nations Secretary-General established, at the request of the General Assembly, five Groups of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security. In 2018, Member States agreed to establish a sixth limited membership Group of Governmental Experts as well as an Open-ended Working Group. See UNGA (2018); UNODA (2020a).

sures to increase trust and stability among states.[41]

In this context, it is reasonable to assume that a single multilateral instrument regulating the military use of AI (e.g. a treaty, a convention) will not be negotiated in the foreseeable future. This does not mean that existing instruments and commitments could not be leveraged to introduce restrictions on military AI applications. For example, prominent among the most widely supported international treaties dealing with the means and methods of warfare are the Geneva Conventions of 1949. More specifically, Article 36 of the 1997 Additional Protocol I to these Conventions (with 177 states parties) may be particularly relevant. The International Committee of the Red Cross, the custodian of the Geneva Conventions and international humanitarian law, describes its function as follows:

> *Article 36 of the 1977 Additional Protocol I to the Geneva Conventions of 1949 requires each State Party to ensure that the use of any new weapons, means or methods of warfare that it studies, develops, acquires or adopts comply with the rules of international humanitarian law. However, all States have an interest in assessing the legality of new weapons, whether or not they are party to Additional Protocol I. Such assessments will contribute to ensuring that the State's armed forces can conduct hostilities in accordance with that State's international obligations.* [42]

While the International Committee of the Red Cross notes that it is up to each state party to establish its own review mechanism, there is an obligation that these mechanisms should constitute a permanent and mandatory procedure for all arms development or acquisition. Although no timetable is specified for the review, the need to conduct it at an early stage of the development of any weapon system is implicit in the obligation. Given this existing commitment, it should be practical to elaborate a measure relevant to military applications of AI to provide some further guidance. For example, states could agree to incorporate AI-enabled decision support systems in the scope of "means or methods of warfare", subjecting such systems to the requirement of legal review.

However, the combination of the secrecy that often characterizes legal reviews with the new, well-documented challenges posed by digital technologies makes the task of conducting legal reviews of AI-enabled military decision support systems particularly challenging. While legal scholars have already started identifying elements or properties for states to consider when conducting Article 36 reviews of AI-embedded means or methods of warfare,[43] significant advances could be made if states with higher capabilities and more experience in this field could share their good practices and lessons learned with other states (see section 4.4.1).

Relatedly, processes for testing, evaluation, validation and verification all are significantly different for AI-based technologies than for traditional weapons. These systems exhibit greater context sensitivity and higher chances of surprising interactions, and so traditional notions of reliability are not always useful. Moreover, methods used in non-military AI contexts often rely on large amounts of field testing (e.g. autonomous vehicles on urban roadways with human oversight) and so are not necessarily usable for military decision support AI systems. Some military organizations have recognized the need to update their testing practices,[44] but multi-stakeholder collaboration (both between states and between states and industry) needs to increase on this issue.

41 See UNGA (2015b).
42 ICRC (2011).
43 See, for example, Boulanin & Verbruggen (2017); Lewis (2019).
44 For example, the US Department of Defense has recently established a subcommittee to address these concerns. See JAIC (2019).

BOX 2 The Challenges of Verification

Agreements are not always adhered to, and the history of arms control and disarmament accords has witnessed numerous occasions when states parties were challenged over their compliance. And in the absence of a supreme authority to arbitrate compliance disputes, real or perceived violations of commitments can lead to the termination of accords (e.g. Russian–US disputes over compliance with the 1987 Intermediate-Range Nuclear Forces Treaty, which resulted in the termination of the agreement in 2019).
As such, a prerequisite for monitoring compliance is the ability to verify good, or bad, conduct by relevant actors (e.g. states, industry).

Challenges of verification, whether technical or political, exist for most types of legally binding arms control instrument but become even more apparent when focusing on intangible technologies such as artificial intelligence: for example, how can verification be conducted on decision-making systems operated by algorithms that do not meet any internationally agreed standard for explainability and predictability?
While these questions about verification might be new in the context of AI, they are not new in the context of arms control. When dealing with tangible, physical items (e.g. missiles, raw materials), accountancy verification methods can be applied (e.g. access control, transfer monitoring, inventory checks).

Conducting verification of intangible technologies is a much more complicated endeavour. Examples exist of efforts by the international arms control community to conduct such verification in both the chemical and biological fields. In most cases, technical means of verification and fact-finding need to be based on strategies and tools more akin to risk assessment and statistically meaningful spot checks than to rigid accountancy controls, such as the system used in nuclear safeguarding.[1]

In the context of the Biological and Toxin Weapons Convention, the Third Review Conference established an Ad Hoc Group of Governmental Experts "to identify and examine potential verification measures from a scientific and technical standpoint".[2] The group, known as VEREX, held four annual sessions in which it examined and evaluated 21 potential verification measures, acknowledging that "some measure in combination could provide enhanced capabilities by increasing, for example, the focus and improving the quality of information, thereby improving the possibility of differentiating between prohibited and permitted activities and of resolving ambiguities about compliance."[3]

Thus, historical examples from other fields of arms control demonstrate that verification, while challenging, is not impossible to achieve when rooted on well-established multilateral, legally binding instruments. In the context of AI, and more specifically AI-enabled decision-making systems, the nature of the technology combined with the absence of any form of international legally binding instrument makes verification difficult to design and implement.

1 Trapp (2019).
2 UNOG (1993, 1).
3 UNOG (1993, 8).

4.2 CONTROLS

4.2.1 Export controls and trade restrictions

Export controls and trade restrictions have traditionally been the most immediate and practical tools to support implementation of non-proliferation-focused arms control by prohibiting or limiting the type or quantity of specific equipment or materials, or restricting who might import them.[45] However, dealing with digital technologies such as AI raises the challenge to a whole different level: What should be restricted or prohibited? How could such restrictions be monitored and enforced?

The Wassenaar Arrangement, a multilateral export control regime governing the export of sensitive dual-use materials, illustrates the challenges of regulating digital technologies using traditional arms control tools.[46] In 2013, the Arrangement's control lists were updated to include intrusion software, in part owing to the implications of such software for human rights (and the risk of abuse). However, the changes, despite being well intentioned, had unintended consequences on cybersecurity research and international collaboration.[47] Many voices raised concerns about the danger of overly broad definitions capturing legitimate uses as well, such as penetration testing tools, which help identify vulnerabilities in computer systems so that they can be patched. In December 2019, members agreed to increased controls over software designed or modified for the conduct of offensive military cyber operations, despite vocal concerns from the private sector and technical community that such controls could negatively impact legitimate cyber vulnerability assessments.[48]

Efforts to adapt dual-use export control regimes to account for the challenges of digital technologies exist also at the regional level. For example, the European Union is currently reviewing its dual-use export control legislation, taking into consideration recent developments and trends, including "rapid scientific and technological developments (e.g. cloud computing and 3-D printing), massive global data networks that are vulnerable to attacks, and the growing availability of cyber tools and information and communications technologies (ICTs) that can be used in violation of human rights".[49]

Countries are also considering these questions at the national level. For example, in the United States of America, a range of AI technologies (e.g. AI chipsets, AI cloud technologies) and applications (e.g. audio and video manipulation technologies, decision support, teaching) have been included in the list of emerging technologies to be subject to the US Export Control Reform Act, which enables increased controls on emerging and foundational technologies deemed essential for national security. The Act supplements the already existing export control framework, which includes the Export Administration Regulations for "dual-use and less sensitive military items" and the International Traffic in Arms Regulations for those articles and services with explicit defence purposes.[50]

One last consideration related to export controls for AI is the need to carefully consider what to control, balancing security needs with commercial considerations. For example, a recent paper by the Center for Security and Emerging Technology argues, "Equipment for manufacturing AI chips is likely a highly effective point of export control. Controls on such equipment effectively constrain who will be able to produce cutting-edge AI chips in the future". It continues, "AI chips themselves are not yet a promising target for expanded regulation. Export controls on AI chips without prior imposition of export controls on the

45 For example, the voluntary Missile Technology Control Regime calls for its members, which include most of the world's key missile manufacturers, to restrict their exports of missiles and related technologies capable of carrying a 500-kilogram payload at least 300 kilometers or delivering any type of weapon of mass destruction. See Davenport (2017).
46 Zetter (2015).
47 Cryptography has been subject to export controls for decades, long before gaining popular attention as the "crypto wars" of the late 1990s, when US export regulations required export licenses for encryption key larger than 40 bits.
48 For a historical overview of the debate surrounding software export controls, see Ruohonen & Kimppa (2019).
49 Immenkamp (2019).
50 Lazarou & Lokker (2019).

equipment for manufacturing such chips will likely prompt targeted countries to invest in chip manufacturing capacity, achieve import substitution…"[51]

4.2.2 Other national controls

In addition to export controls that focus on granting or denying access to a given technology, national controls could also be put in place focusing on AI in military decision-making processes, for example through the development of a dedicated national directive.[52]

National directives are transparent, authoritative means through which governments can set the bar for their own internal stakeholders (e.g. different parts of the national security apparatus), for the research and development community, and for industry, as well as, in some cases, influence policy development by other governments.

The content of such directives should be aligned with their purpose: creating a common baseline understanding of concepts, actors, roles and responsibilities for the use of AI in military applications, including those supporting decision-making. This would include, for example:

» Providing clear and shared definitions of concepts

» Assigning clear levels of responsibility to different actors throughout the life cycle of any given AI military application (e.g. from development, to testing, fielding, and employment)

» Describing the overarching policy on the use of such AI applications

» Setting clear conditions and overarching requirements that any AI military application would need to satisfy (e.g. the need to ensure appropriate levels of human involvement)

» Providing guidelines for specific issues (e.g. testing and evaluation of such applications)

As an illustrative example, the US Department of Defense Directive 3000.09 "Autonomy in Weapon Systems", through its main text and supporting appendices, covers issues such as purpose, applicability, definitions, policy, responsibilities, guidelines for review, verification, validation, testing and evaluation of autonomy in weapon systems.[53] As the first such publicly available national directive on autonomy, the directive was widely discussed in the framework of the CCW discussion on autonomous weapons and has been influential in shaping the terms of those talks and the positions of some other states.

The international community, within the framework of the CCW, could consider developing a model of such a national directive to be used as a blueprint for states desiring guidance on developing their own.[54] Examples of when this has been done in the past include the development of model laws to support the implementation of the United Nations Firearms Protocol[55] or, at the regional level, the assistance in developing draft model legislation to help Caribbean Community member states implement the Arms Trade Treaty.[56]

The development of directives through a genuinely multi-stakeholder approach, where different interested parties (from all sectors) are invited to join and contribute, can help communities come together and break down silos, encouraging diversity of perspectives.

51 Flynn (2020).
52 US DoD (2017).
53 US DoD (2017).
54 In 2021, the CCW will undergo its Sixth Review Conference, during which the future of the Group of Governmental Experts on Lethal Autonomous Weapons Systems will also be discussed. Should the High Contracting Parties decide to renew the Group of Governmental Experts for one or more years, the development of such a model directive could be included in its mandate.
55 UNODC (2011).
56 ATT Assistance (2016).

Once launched, a national directive on the use of AI in military decision-making could have a significant impact across a wide range of stakeholders: industry, contractors, research and technical communities, and allies and partners (including military alliances). However, it would be important that such a directive include a periodic review process (e.g. every five years) calling for a rigorous evaluation and impact assessment. This is key to ensuring that, at the very least, the directive (1) remains relevant in light of technological developments; (2) is achieving the desired objectives; and (3) does not create unacceptable, unanticipated effects. An example of a methodology for such periodic assessments is provided by the European Union's Better Regulation Guidelines.[57]

BOX 3 Possible Approaches to Technical Controls

Export controls would target access to artificial intelligent (AI) technology. Directives, strategies and policies would target the use of AI technology. A last category of controls is technical measures, which could be used to target the behaviour of the technology itself.

Recall mechanisms, similar to those used in the automotive and software industries, would require technology suppliers to promptly recall applications for which a bug or vulnerability has been identified and block its use until a patch has been installed. One of the main challenges of this type of measure is that AI applications continuously evolve as they feed their machine learning algorithms with new data. To be able to promptly detect unintended behaviours of the application, a technology supplier would have to require continuous monitoring of its use, which might be problematic in a military context.

Built-in switches could be used for a variety of purposes, from simple "auto-update" modes, to a "return to base" or "abort mission" mode for system that manages units deployed on the battlefield, to forced deactivation. This kind of switch could be particularly useful in the context of a dynamic environment, where the situation and the data are constantly evolving. Some of these switches could be automated using simple rule-based models (e.g. If This Then That). The limitations of such automated switches are that they assume the system can detect the internal trigger (e.g. failure of one of the sensors) or external trigger (e.g. change of environment). As this may not always be the case, human operators should also have the ability to activate such switches.

This approach would require a considerable investment in research and development and would be dependent on military authorities being willing to impose certain basic restrictions on military AI applications. However, such a technical remedy would offer the benefits of an automated control that would not be influenced by subjective judgments in stressful circumstances.

57 The Better Regulation Guidelines set out the principles that the European Commission follows when preparing new initiatives and proposals and when managing and evaluating existing legislation. For more information, see European Commission (2019).

4.3 VOLUNTARY MEASURES

4.3.1 Norms

Norms and standards are potentially powerful ways to promote the responsible and safe development, adoption and use of a technology.

In 2015, the General Assembly agreed by consensus[58] that all states should be guided by 11 voluntary norms of responsible state behaviour in their use of ICTs in the context of international security.[59] To date, these norms represent the highest multilateral accomplishment in the area of digital technologies and their implications for international security. In recent years, additional norm-setting initiatives on cybersecurity have been launched under the leadership of the private sector (e.g. the Cybersecurity Tech Accord or the Charter of Trust) or through multi-stakeholder partnerships (e.g. the Paris Call).

In the field of AI, although different sets of principles have been developed by different actors (see section 4.3.3), no equivalent multilaterally agreed norms exist for AI. In recent years, states have considered AI-enabled physical weapon systems in discussions on autonomous weapon systems.[60] At the heart of these discussions have been definitional issues and consideration of the legal and ethical issues that arise when delegating an increasing number of military tasks to machines. The most significant achievement so far has been the endorsement by the High Contracting Parties to the CCW in 2019 of 11 guiding principles.[61] While short of being fully developed norms, these guiding principles – by virtue of having been negotiated and adopted at the multilateral level – represent a solid foundation for further developments, potentially branching out of the realm of physical weapon systems to encompass wider military applications of AI, including those that support decision-making.

4.3.2 Standards

Another area in which the transferability of good practices from the cyber and ICT sector could potentially be explored relates to the development of standards.

A large body of standards exists to regulate all aspects of cyber and ICT security, from standards set by national agencies (e.g. the US National Institute of Standards and Technology)[62] to standards developed at the international level (e.g. the International Telecommunication Union,[63] the International Organization for Standardization [ISO][64] and the International Electrotechnical Commission[65]).

In the field of AI, the development of standards for attributes such as predictability and explainability would be particularly useful for military decision support AI systems.[66] In a crisis, the ability to refer back to an agreed set of definitions and technology standards can be a helpful tool to rapidly establish a common understanding of the situation: for example, a statement such as "the AI running on system X is compliant with explainability standard Y" would include in its essence a significant amount of information that would be publicly known and recognized as acceptable by a national authority, a group of like-minded states or an international standardization body.

Although such efforts would naturally start with voluntary mechanisms, it does not mean that they could not become highly impactful. For example, ISO standards began as a voluntary stamp of approval, but over time in certain countries they have become a legal or commercial requirement, shaping laws, procurement processes and business operations. Numerous efforts are

58 UNGA (2015a).
59 As recommended in UNGA (2015b).
60 UNOG (2020b).
61 Group of Governmental Experts (2019).
62 NIST (2020).
63 ITU (2020).
64 ISO (2020).
65 IEC (2020).
66 Cihon (2019).

BOX 4 A possible approach to standardizing the certification process for military applications of artificial intelligence

A key step in the process that leads to the integration of a new technology and its subsequent fielding is certification. When it comes to artificial intelligence (AI) applications in military decision-making, such a certification process could be based on, for example, the following five key features:

- Early **engagement with end-users** and those expected to engage with the application (e.g. downstream decision makers or operators). This is key to ensuring that the technology is addressing an actual need or gap (e.g. enabling something that was not possible before or improving the ways in which current tasks are performed).
- **Adversarial testing**. Performing adversarial testing through red-teaming methods is key to assessing the behaviour of the application in contexts that are, by definition, hostile and dynamic.
- Proper **training and education**. It is fundamental that proper training and education occurs, not only of the end-users but of the entire chain of military command that determines if, when and how to deploy such applications. This is key to fully understanding the capabilities and limitations of each application and limiting the effects of cognitive bias (making decisions based on what a human believes the technology can do instead of on a well-developed understanding of real capabilities).
- Clear specification of, and notification about, the **intended and permissible uses** for the decision support system. Certifications should include "acceptable use terms" – similar to those already included in software contracts – for AI-powered decision support systems. This would also feed into any associated legal review processes of such systems.
- **Continuous certification**. For AI-based applications, certification should be considered as a dynamic and continuous process that accounts for changes in the technology and in the environment in which it is deployed.

under way to develop AI-specific ISO standards or to adapt existing ISO standards to the use of AI systems. Those efforts are not yet sufficiently refined or widely adopted to provide a basis for the certification of military decision support AI systems. Much more work remains to be done.

4.3.3 Codes of conduct and principles

A literal definition of **code of conduct** is "a collection of rules and regulations that include what is and is not acceptable or expected behavior".[67] Codes of conduct have been used throughout the centuries as "soft regulation" by different types of actor, from social groupings, to professional orders, to businesses, to states. **Principles** often provide the basis or foundation for codes of conduct.

When applied to the field of international security and arms control, codes of conduct are the bridge between voluntary measures and CBMs (see section 4.4). An example of this category is The Hague Code of Conduct, which regulates the area of ballistic missiles capable of carrying weapons of mass destruction. Its 135 current members have made a voluntary political commitment to provide pre-launch notifications on ballistic missile and space-launch vehicle launches and test flights, as well as to submit an annual declaration of their country's policies on ballistic missiles and space-launch vehicles.[68]

67 YourDictionary (2020).
68 The HCOC (2020).

In the field of AI, many initiatives led by industry, national governments or international organizations have approached "soft regulation" by means of AI principles. A few examples include

» Industry:
  - Google's AI principles,[69] Microsoft's AI principles,[70] Philips's AI principles[71]

» National agencies:
  - Government of the United Kingdom, code of conduct on artificial intelligence systems used by the National Health Service[72]
  - Government of Australia, Department of Industry, Science, Energy and Resources, AI principles[73]
  - Monetary Authority of Singapore, *Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector*[74]

» International organizations:
  - Organisation for Economic Co-operation and Development, Principles on Artificial Intelligence[75]
  - IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems[76]
  - Association for Computing Machinery, ACM Code of Ethics and Professional Conduct[77]

While different in scope (e.g. from self-regulation of industry to agreements among members of an international organization), many of these sets of principles are similar in nature as they cover issues such as transparency, reliability, fairness and accountability.[78] There is also recognition that principles need to be operationalized. For example, in February 2020, a year after the release of its AI strategy, the US Department of Defense adopted a set of principles on the ethical use of AI.[79] Using these principles, the Department will work on implementation guidelines on issues such as procurement, safeguards, risk mitigation and training.[80]

Leveraging these commitments and expressions of good intentions at the multilateral level remains an underused tool in international security discussions. As principles and codes are operationalized, there is an opportunity for sharing good practice and lessons learned, which is a traditional CBM.

4.4 CONFIDENCE-BUILDING MEASURES

4.4.1 Building trust, transparency and confidence

CBMs in the field of arms control are voluntary measures designed to prevent hostilities, avert escalation, reduce military tension and build mutual trust between countries or communities.[81]

CBMs can be unilateral, bilateral or multilateral and take different forms depending on the specific context in which they are applied (e.g. pre- or post-conflict, intra- or inter-state). Although they have been applied for centuries on all continents, the first codified CBMs appeared only in the second half of the twentieth century as initial steps to increase transparency during the Cold War.[82] Two examples of

69 Google AI (2020).
70 Microsoft (2020).
71 Philips (2020).
72 Department of Health and Social Care (2019).
73 Australian Government, Department of Industry, Science, Energy and Resources (2019).
74 Monetary Authority of Singapore (2018).
75 OECD (2019).
76 IEEE (2020).
77 ACM (2020).
78 For a useful analysis of consensus among these documents, see Fjeld & Nagy (2020).
79 US DoD (2020).
80 Williams (2020).
81 UNODA (2020b).
82 OSCE (2012, 11).

significant milestones achieved in modern times are the 1986 Stockholm Document,[83] which included the first set of militarily and politically binding verifiable CBMs, and the 1990 Vienna Document,[84] which includes a series of CBMs covering both immediate risk reduction and long-term routine military interaction (e.g. on-site inspections and evaluation visits, annual exchanges of military information and dialogue on defence planning).[85] At the United Nations level, an important milestone is the United Nations Disarmament Commission's 1988 *Guidelines for appropriate types of confidence-building measures and for the implementation of such measures on a global or regional level*,[86] complemented by a more recent set of recommendations on practical CBMs in the field of conventional weapons.[87] Although not all traditional military CBMs can be transferred to the digital world, much has been done over the past decade to find valid alternatives that could support a peaceful and safe cyberspace. For example, CBMs have been a key component of multilateral negotiations on cybersecurity under United Nations auspices.[88]

In the context of developing and implementing CBMs, regional organizations have a key role to play, given their nuanced understanding of the local context and more immediate access to relevant stakeholders. In the context of cyber, the Organization of American States, the Organization for Security and Co-operation in Europe (OSCE) and the Association of Southeast Asian Nations are three good examples of regional organizations taking CBMs on digital technologies forward in ways that are specifically relevant to the needs and concerns of their members. These CBMs cover issues such as exchanging information and views on threats to and in the use of ICTs; holding consultations to reduce the risk of misperception; disseminating best practices; nominating national points of contact at the policy level; and sharing information on their national organization, strategies, policies and programmes relevant to the security and use of ICTs, including in public-private partnerships.[89]

83 OSCE (1986).
84 OSCE (1990).
85 OSCE (2012, 12).
86 UNGA (1999b).
87 UNGA (2017).
88 The Sixth Group of Governmental Experts and the more recently established Open-ended Working Group. See UNODA (2020a).
89 See, for example, CICTE (2017); OAS General Assembly (2018); OSCE (2013); Zannier (2014).

BOX 5 Characteristics of successful confidence-building measures

In designing confidence-building measures for artificial intelligence-enabled military decision support, it is key to remember that while CBMs developed for information and communications technology can be a reference point, CBMs must be tailored to the specific context. That being said, the common characteristics shared by successful CBMs, as proposed by the Organization for Security and Co-operation in Europe, is applicable even in the context of AI. These characteristics include:

1. **Reciprocal**: Measures taken by one party should lead (not necessarily immediately) to similar measures being undertaken by the other party (or parties) in a balanced and reciprocal manner, in accordance with the principle of mutual benefit.
2. **Incremental**: Starting with small, less ambitious measures can support the building of trust necessary to move toward more complex and difficult measures.
3. **Long term oriented**: Building confidence takes time, and for CBMs to be effective they need to be long-term commitments to avoid setbacks.
4. **Predictable**: The nature, scope and content of CBMs should be predictable and promote predictable behaviours.
5. **Transparent**: The intent and modalities of CBMs should be obvious and unambiguous.
6. **Reliable**: CBMs should be reliable, meaning, for example, that they should not be used as political tactical manoeuvres.
7. **Consistent**: CBMs should be consistent with target groups, their topics and the message they send.
8. **Verifiable**: Verification of CBMs, particularly those where reciprocity is expected, is an important component of trust and confidence building.
9. **Locally owned**: The long-term success of CBMs relies on the level of engagement and commitment of the targeted groups or actors.
10. **Multilevel or multi-stakeholder**: For CBMs to be successful, governments must mobilize and engage with the broader society.
11. **Supported by appropriate communication channels**: Efficient communication and information flow is a key enabler of successful CBMs to address misunderstandings in a timely manner.

Source: OSCE (2012, 16–18).

4.4.2 CBMs for AI-enabled military decision support

The development of national directives or of national and international standards is a measure that indirectly contributes to creating a transparent environment for states, industry, academia, civil society and the general public.

Information-sharing enables transparency which, in turn, helps enable trust. However, as different stakeholders play distinct roles within the AI ecosystem, it is important to differentiate between types of measure that could be designed for different types of interaction (e.g. government-to-government; between government and industry; between government, industry and the wider scientific community).

A measure that would support transparency and trust across all types of interaction is the development and public release of a national regulatory framework for AI-enabled military decision support. This could include, for example:

» AI strategies, policies and directives

» Guidelines describing verification and validation procedures, testing and evaluation processes, and data management and protection measures

» Details about the processes for conducting legal reviews of AI-enabled military decision support systems (not necessarily the results of the reviews themselves)

Sharing these types of information publicly would achieve a double purpose: support confidence building among states and provide a baseline or blueprint for other governments in the process of developing their own regulatory frameworks.

GOVERNMENT TO GOVERNMENT

When focusing at the **government-to-government** level, information-sharing can take different forms. An example of where information-sharing is working well is the Biological and Toxin Weapons Convention, where states parties agreed at the Second Review Conference to the exchange of CBMs "in order to prevent or reduce the occurrence of ambiguities, doubts and suspicions and in order to improve international cooperation in the field of peaceful biological activities".[90]

In the context of military applications of AI for decision support, in the absence of an overarching multilateral framework (e.g. a dedicated international convention) it is difficult to institutionalize information-sharing and other CBMs. In the absence of such a formal multilateral anchor, government-to-government engagements currently remain ad hoc through the use of other channels and platforms, for example at the margins of multi-stakeholder initiatives such as the Global Partnership on Artificial Intelligence, advanced by Canada and France, or through dedicated bilateral dialogues.

Although military applications of AI are considered highly sensitive, it is important that these engagements continue on all diplomatic tracks, even if they remain at a high level of generality and fail to address specific regulatory requirements. The added value of such engagements often resides more in the building of the networks than in the actual substance discussed. In this context, **middle powers and industry groups (or civil society organizations)** can play an important role as conveners and mediators of such engagements among major military powers.

GOVERNMENT TO INDUSTRY

With regard to **industry**, it is key that governments provide as much guidance as possible on their requirements (through a regulatory framework, as described earlier in this chapter). That being said, as mentioned in section 3.2, it is important to recall that whereas historically much of the relevant non-governmental scientific expertise was in academia, in AI a large part is in the private sector. As such, even in the absence of such a regulatory framework, increased transparency from industry concerning its own practices and procedures for technology development, verification, testing and integration, as well as on overall capabilities, limitations and optimal use of the system, would help raise standards. The interface between industry and government is likely to be the area that benefits most from the **development of standards**, to the extent that standards can support the creation of a shared understanding of the technology, its performance and its limitations.

MULTI-STAKEHOLDER DIALOGUE AND ENGAGEMENT WITH THE WIDER SCIENTIFIC COMMUNITY

An alternative approach to overcoming the sensitivities of engaging in transparency and trust-building initiatives would be to leverage a bottom-up approach based on **multi-stakeholder or industry-led initiatives and exchanges between scientific communities** before an application enters commercial development or is fully integrated into a military capability.

This more peripheral layer of CBMs (as opposed to those centred on governments) is particularly relevant for two reasons:

» It includes the vast majority of technical and scientific knowledge on this issue.

90 Second Review Conference (1986).

» It is potentially less impacted by an adverse political environment.

The advantages of such an approach are that the academic and scientific communities enable exchanges and collaborations by design, and their members are naturally inclined to share information, solicit feedback and advance knowledge. If properly channelled, such a collective body of knowledge could be used to stimulate upstream transparency and trust-building. For example, increased exchanges among developers, researchers and scientists could remain at a completely unclassified level but be focused on the risks of specific military applications or use cases, generating a body of valuable knowledge.

Specific activities could include dedicated academic conferences or the organization of prizes or award-based competitions, such as the UK Ministry of Defence's "Grand Challenges", with a point-based system that could incentivize international cooperation. Some have suggested the establishment of a neutral international scientific research centre on AI, similar to CERN, where Track 2 "science diplomacy" could build on the open culture of scientific research and development as well as the socialization of norms and best practices, particularly on AI safety.[91]

Engagement with students and young scientists also offers a powerful entry point for establishing a culture of engagement on international security with the research community. For example, the International Genetically Engineered Machine competition is an annual global team-based competition for undergraduates in synthetic biology to design, build, test and measure a system. As part of their submission, teams consider safety and security aspects of their design, which in the process raises awareness of not only responsible innovation but also the provisions of the biological weapons regime. In a similar fashion, AI challenges might systematically include a component that considers the implications of specific innovations or applications for international security.

All the initiatives described above could be reinforced by the organization of regular multi-stakeholder dialogues that bring together representatives of different communities, enabling better understanding of one another's work and offering opportunities for the creation of cross-sector networks and working relationships.

4.5 THE ROLE OF INCENTIVES

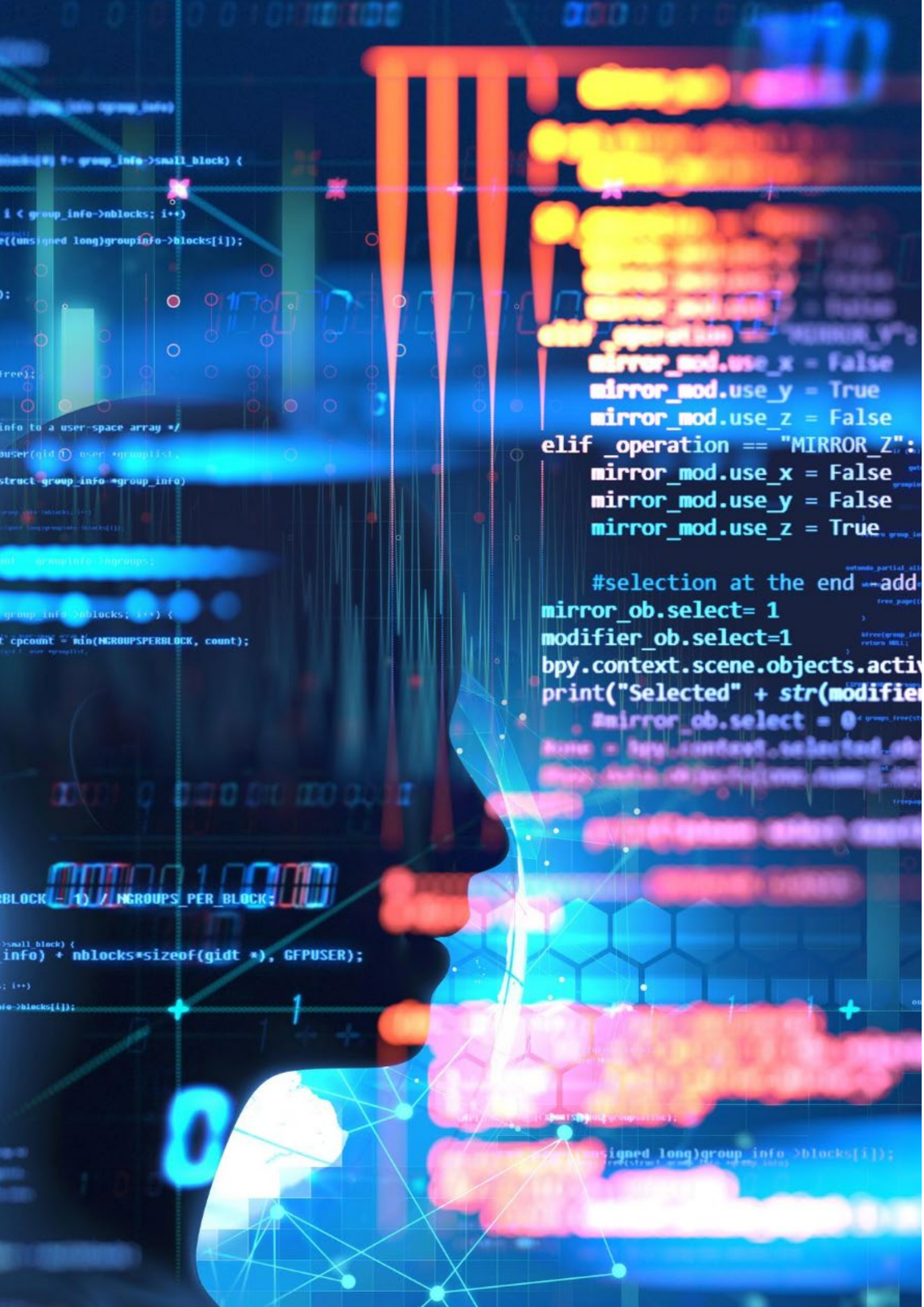
An important consideration related to the development and subsequent adoption of voluntary measures (norms and standards) or CBMs for AI applications in military decision support systems is related to incentives: What incentives do states and other stakeholders have to develop and adopt such norms and standards? Why should government, industry and, to some extent, academia invest time and resources in, for example, information-sharing?

The current narrative describes the military interest in AI as an "AI arms race". This connotation of the current landscape incentivizes a closed, opaque and competitive culture aimed at "winning the race", which may ultimately lead to a "race to the bottom", with the premature – and unsafe – deployment of AI applications to leverage a first mover advantage.

If, instead, the narrative could pivot to a more positive and constructive aim, such as achieving a more reliable, better integrated and more easily explainable technology, then incentives would be created for states and industry alike to engage in the development of such standards, independently from the intended military use of the full capability once developed. Recognizing that the current international context may hamper a more ambitious multilateral approach in this field, these efforts could begin with one or a small group of like-minded states taking the lead in developing and adopting norms and standards on a voluntary basis and, in doing so, establishing a baseline of good or best practice for others to follow. This would showcase thought leadership and could be a salient norm-setting power. Norms, standards and CBMs are also important components of a

91 See, for example, Fischer and Wenger (2019).

risk reduction approach that aims to avoid a technological arms race based on wrong assumptions about what others are doing and to reduce the risk of unintended consequences of systems that are developed in isolation and react when they come into contact with one another in a way that was never foreseen or intended (or desirable).


mirror_mod.use_x = False
mirror_mod.use_y = True
mirror_mod.use_z = False
elif _operation == "MIRROR_Z":
mirror_mod.use_x = False
mirror_mod.use_y = False
mirror_mod.use_z = True
#selection at the end
mirror_ob.select= 1
modifier_ob.select=1
print("Selected" + str(modifier
#mirror_ob.select = 0
struct group_info *group_info)
group_info->nblocks; i++) {
t cpcount = min(NGROUPSPERBLOCK, count);
BLOCK - 1) / NGROUPS_PER_BLOCK;
info) + nblocks*sizeof(gidt *), GFPUSER);
fo->blocks[i]);
info to a user-space array */
i < group_info->nblocks; i++)
((unsigned long)groupinfo->blocks[i]);
free);
group_info->small_block) {

5. THE WAY FORWARD

This report has deliberately focused on an area of military AI that, while having the potential to undermine security and stability, is not currently discussed within any multilateral process: the use by militaries of AI-enabled decision support systems. Looking at the relevance of arms control tools through the lens of this specific use case of AI highlights that although many of the traditional tools of arms control remain as relevant as ever, new ways of working and new relationships will be necessary to address these challenges effectively.

A number of key takeaways can be extracted from this report, including some that might be relevant for multilateral discussions that go beyond the specific application of AI to military decision support and include wider applications of digital technologies (from cyber to autonomous weapon systems). In particular:

1. **AI in military decision support systems is an area that deserves further attention** from the international security community in order to manage risk and potential for instability.

2. The traditional **objectives of arms control remain valid and applicable** even when dealing with AI-enabled decision support systems.

3. The traditional arms control toolbox will not become obsolete if the arms control community is open and willing to **embrace new forms of collaboration as well as adapt traditional ones** to fully **leverage the know-how of the scientific expert community, most of which now resides in the private sector**. This also entails creating more opportunities for exchanges and cross-fertilization of ideas and perspectives by involving industry and scientific experts in relevant arms control discussions and by ensuring that the arms control community actively engages with such actors in their relevant forums.

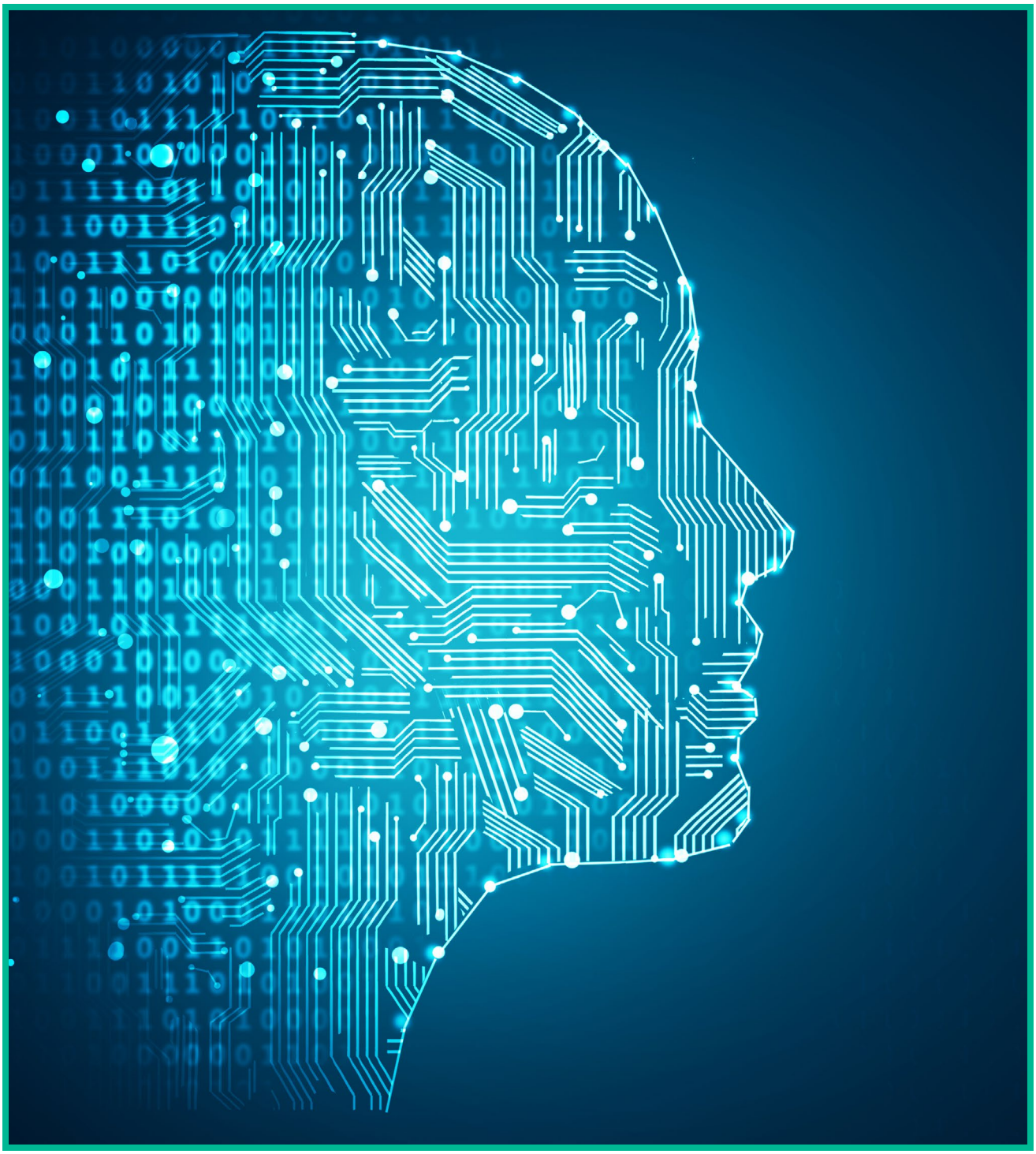
4. **There is no "one stop" solution**. A web of responses and incentive structures that target and draw on the expertise of different stakeholder groups will be required to effectively respond to the challenges posed by embedding AI in decision support applications.

5. While governments remain the natural owners of traditional arms control tools, the **range of possible measures** described in this report **does not always require government leadership** nor, in some cases, government direct participation (e.g. industry-led standardization processes, scientific knowledge exchanges). However, for these measures to produce meaningful impact on strategic stability and security, they would require recognition and downstream support by state actors. **Industry** can play a critical role provided it is given the opportunity to meaningfully engage with the arms control community.

6. Building on this last point, as a thought-leader in AI, **industry has its own responsibilities**: from including legal and ethical considerations in their innovation policies and practices applied to AI, to being willing to consistently engage with regulatory processes at the international and national levels. This is particularly relevant for industry developing AI directly for defence, or for the broader industrial base developing potential dual-use AI or applications. Current efforts by a range of private sector actors to develop standards, or principles of responsible development of AI (e.g. transparency, reliability, security) are particularly relevant for the current debate on the international security implications of AI.

This report provides an initial insight into why the international security community may need to consider regulating AI applications that fall in the digital grey zone between AI-enabled weapon systems (e.g. lethal autonomous weapon systems) and military uses of civilian AI applications (e.g. logistics, transport), as well as an initial exploration of the variety of familiar tools the community has at its disposal to do so.

While the arms control community has been focused for the past several years on autonomous weapon systems, wider consideration of the international security dimension of AI-enabled decision support tools may logically flow from the community's deliberations. Additional research on the international security implications of technical aspects of algorithmic decision-making (e.g. explainability, predictability) in both weapon and decision support tools, and further exploration of softer regulatory approaches with greater involvement from the research and technical communities, as well as industry, may be a practical contribution to this endeavour.



BIBLIOGRAPHY

ACM (Association for Computing Machinery). 2020. 'ACM Code of Ethics and Professional Conduct'. As of 29 May: https://www.acm.org/code-of-ethics

Albright, Alex. 2019. 'If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions'. The John M. Olin Center for Law, Economics, and Business Fellows' Discussion Paper Series 85. As of 27 July 2020: https://thelittledataset.com/about_files/albright_judge_score.pdf

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. 'Concrete Problems in AI Safety'. arXiv:1606.06565v2 [cs.AI]. As of 27 July 2020: https://arxiv.org/pdf/1606.06565.pdf

Apps, Peter. 2019. 'COLUMN-Commentary: Are China, Russia Winning the AI Arms Race?' Reuters, 15 January. As of 27 July 2020: https://uk.reuters.com/article/apps-ai/column-commentary-are-china-russia-winning-the-ai-arms-race-idUKL1N1ZF23D

Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. 'Synthesizing Robust Adversarial Examples'. In *Proceedings of the 35th International Conference on Machine Learning*, edited by Jennifer Dye & Andreas Krause, 449–468. As of 27 July 2020: https://arxiv.org/pdf/1707.07397.pdf

ATT (Arms Trade Treaty) Assistance. 2016. 'CARICOM Ongoing Project to Develop Draft Model Legislation for ATT Implementation: Sipri'. As of 19 May 2020: http://www.att-assistance.org/activity/caricom-ongoing-project-develop-draft-model-legislation-att-implementation

Atwood, David C. 1999. 'Implementing Ottawa: Continuity and Change in the Roles of NGOs'. *Disarmament Forum 4*: 19–32. https://unidir.org/files/publications/pdfs/framework-for-a-mine-free-world-en-367.pdf

Australian Government, Department of Industry, Science, Energy and Resources. 2019. 'AI Ethics Principles'. As of 2 September 2019: https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles

Barocas, Solon, & Andrew Selbst. 2016. 'Big Data's Disparate Impact'. *California Law Review*, 104: 671–732. As of 27 July 2020: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899

Borrie, John. 2019. 'Cold War Lessons for Automation in Nuclear Weapon Systems', In *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk, Vol. I. Euro-Atlantic Perspectives*, edited by Vincent Boulanin, 41–52. Stockholm: Stockholm International Peace Research Institute. As of 27 July 2020: https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf

Boulanin, Vincent, & Maaike Verbruggen. 2017. *Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies*. Stockholm: Stockholm International Peace Research Institute.

Campaign to Stop Killer Robots. 2020. 'About Us'. As of 29 May: https://www.stopkillerrobots.org/about

CICTE (Inter-American Committee Against Terrorism). 2017. *Establishment of a Working Group on Cooperation and Confidence-Building Measures in Cyberspace*, CICTE/RES.1/17, 10 April 2017.

Cihon, Peter. 2019. *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. Oxford: Future of Humanity Institute, University of Oxford. As of 27 July 2020: https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf

Cohen, Zachary. 2017. 'US Risks Losing AI Arms Race to China and Russia.' CNN.com, 29 November, 13.57 GMT. As of 27 July 2020: https://edition.cnn.com/2017/11/29/politics/us-military-artificial-intelligence-russia-china/index.html

Cybersecurity Tech Accord. 2020. 'About the Cybersecurity Tech Accord'. As of 30 April: https://cybertechaccord.org/about

Danks, David. 2020. 'How Adversarial Attacks Could Destabilize Military AI Systems'. IEEE Spectrum, 26 February, 16.51 GMT As of 27 July 2020: https://spectrum.ieee.org/automaton/artificial-intelligence/embedded-ai/adversarial-attacks-and-ai-systems

Danks, David, & Alex John London. 2017. 'Algorithmic Bias in Autonomous Systems'. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence,* edited by C. Sierra, 4691–97. Palo Alto: AAAI Press.

Davenport, Kelsey. 2017. 'The Missile Technology Control Regime at a Glance: Fact Sheets & Briefs'. Arms Control Association. As of 19 May 2020: https://www.armscontrol.org/factsheets/mtcr

Deeks, Ashley. 2018. 'Predicting Enemies'. *Virginia Law Review* 104: 1529–93. As of 27 July 2020: https://www.virginialawreview.org/sites/virginialawreview.org/files/Deeks_Online_0.pdf

Department of Commerce Bureau of Industry and Security. 2018. 'Review of Controls for Certain Emerging Technologies.' *Federal Register* 83(223): 58201, 19 November. As of 27 July 2020: https://www.govinfo.gov/content/pkg/FR-2018-11-19/pdf/2018-25221.pdf

Department of Health and Social Care. 2019. 'New Code of Conduct for Artificial Intelligence (AI) Systems Used by the NHS.' GOV.UK, 19 February. As of 27 July 2020: https://www.gov.uk/government/news/new-code-of-conduct-for-artificial-intelligence-ai-systems-used-by-the-nhs

Dietvorst, Berkley J., Joseph P. Simmons, and Cade Massey. 2015. 'Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err'. *Journal of Experimental Psychology: General* 144(1): 114–26. doi: http://dx.doi.org/10.1037/xge0000033

DoD (US Department of Defense). 2017. *Autonomy in Weapon Systems*. Directive 3000.09, 21 November 2017. As of 27 July 2020: https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf

———. 2020. 'DOD Adopts Ethical Principles for Artificial Intelligence'. As of 3 June 2020: https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence

EMSA (European Maritime Safety Agency). 2020. 'Accident Investigations'. As of 19 May: http://www.emsa.europa.eu/implementation-tasks/accident-investigation.html

European Commission. 2019. 'Better Regulation: Guidelines and Toolbox'. As of 27 July 2020: https://ec.europa.eu/info/law/law-making-process/planning-and-proposing-law/better-regulation-why-and-how/better-regulation-guidelines-and-toolbox_en

European Parliament and Council of the European Union. 2009. *Establishing the fundamental principles governing the investigation of accidents in the maritime transport sector and amending Council Directive 1999/35/EC and Directive 2002/59/EC of the European Parliament and of the Council*. Directive 2009/18/EC, 23 April. As of 27 July 2020: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:131:0114:0127:EN:PDF

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno & Dawn Song. 2018. 'Robust Physical-World Attacks on Deep Learning Visual Classification'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1625–34. New York: As of 27 July 2020: https://arxiv.org/abs/1707.08945

Ferguson, Andrew G. 2017. 'Policing Predictive Policing'. *Washington University Law Review* 94(5): 1109–90. As of 27 July 2020: https://openscholarship.wustl.edu/law_lawreview/vol94/iss5/5

Fischer, Sophie-Charlotte, & Andreas Wenger. 2019. 'A Politically Neutral Hub for Basic AI Research'. *CSS ETH Zurich Policy Perspectives* 7(2). https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/PP7-2_2019-E.pdf

Fjeld, Jessica, & Adam Nagy. 2020. 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI'. Berkman Klein Center for Internet and Society at Harvard University, 15 January. As of 27 July 2020: https://cyber.harvard.edu/publication/2020/principled-ai.

Flynn, Carrick. 2020. 'Recommendations on Export Controls for Artificial Intelligence'. Issue Brief. Washington, DC: Center for Security and Emerging Technology. As of 27 July 2020: https://cset.georgetown.edu/wp-content/uploads/Recommendations-on-Export-Controls-for-Artificial-Intelligence.pdf

Freedberg Jr., Sydney J. 2019. "No AI For Nuclear Command & Control: JAIC's Shanahan." Breaking Defense, 25 September. As of 27 July 2020: https://breakingdefense.com/2019/09/no-ai-for-nuclear-command-control-jaics-shanahan/#:~:text=GEORGETOWN%20UNIVERSITY%3A%20"You%20will%20find,with%20nuclear%20command%20and%20control."

GGE (Group of Governmental Experts) on Lethal Autonomous Weapons Systems. 2018. *Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*. UN document CCW/GGE.1/2018/3, 9–13 April 2018 and 27–31 August 2018. As of 27 July 2020: https://www.unog.ch/80256EDD006B8954/(httpAssets)/20092911F6495FA7C125830E003F9A5B/$file/CCW_GGE.1_2018_3_final.pdf

———. 2019. *Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*. UN document CCW/GGE.1/2019/3, 25–29 March 2019 and 20–21 August 2019. As of 27 July 2020: https://undocs.org/en/CCW/GGE.1/2019/3

Goldblat, Jozef. 2002. *Arms Control: The New Guide to Negotiations and Agreements*. Thousand Oaks: Sage Publications.

Google AI. 2020. 'Artificial Intelligence at Google: Our Principles.' As of 19 May: https://ai.google/principles

Guidotti Ricardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. 'A Survey of Methods for Explaining Black Box Models'. *ACM Computing Surveys* 51(5): 93. doi: https://doi.org/10.1145/3236009

Hammond, G. 1993. *Plowshares into Swords: Arms Races in International Politics, 1840–1991*. Columbia: University of South Carolina Press.

Hughes, Mark. 2017. 'Artificial Intelligence Is Now an Arms Race. What If the Bad Guys Win?' World Economic Forum, 10 November. As of 18 May 2020: https://www.weforum.org/agenda/2017/11/cybersecurity-artificial-intelligence-arms-race

ICAO (International Civil Aviation Organization). 2020. 'Safety'. As of 19 May: https://www.icao.int/safety/Pages/default.aspx

ICRAC (International Committee for Robot Arms Control). 2020. 'About ICRAC'. As of 29 May: https://www.icrac.net/about-icrac

ICRC (International Committee of the Red Cross). 2011. 'Review of New Weapons.' 30 November.

IEC (International Electrotechnical Commission). 2020. 'Cyber Security'. As of 19 May: https://www.iec.ch/cybersecurity

IEEE (Institute of Electrical and Electronics Engineers). 2020. 'The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems'. As of 29 May: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

Immenkamp, Beatrix. 2019. *Review of Dual-Use Export Controls*. Briefing. Brussels: European Parliament Research Service. https://www.europarl.europa.eu/RegData/etudes/BRIE/2016/589832/EPRS_BRI(2016)589832_EN.pdf

ISO (International Organization for Standardization). 2020. 'ISO/IEC 27001 – Information Security Management'. https://www.iso.org/isoiec-27001-information-security.html

ITU (International Telecommunication Union). 2020. 'ICT Security Standards Roadmap'. As of 18 July: https://www.itu.int/en/ITU-T/studygroups/com17/ict/Pages/default.aspx

JAIC (Joint Artificial Intelligence Center). 2019. 'The DoD AI Ethical Principles – Shifting from Principles to Practice'. AI in Defense, 1 April. As of 27 July 2020: https://www.ai.mil/blog_04_01_20-shifting_from_principles_to_practice.html

Johnson, Rebecca. 2011. *Experts, Advocates and Partners: Civil Society and the Conference on Disarmament*. CD Discussion Series. Geneva: UNIDIR. As of 27 July 2020: https://unidir.org/files/publications/pdfs/civil-society-and-the-conference-on-disarmament-360.pdf

Jovanovic, Boyan, & Peter Rousseau. 2005. 'General Purpose Technologies'. In *Handbook of Economic Growth*, Volume 1B, edited by Philippe Aghion & Steven N. Durlauf. Amsterdam: Elsevier.

King's College London and Geneva Disarmament Platform. 2017. 'Civil Society and the BWC: Finding a Way Forward'. Workshop Report. Geneva. As of 27 July 2020: http://www.filippalentzos.com/wp-content/uploads/2018/03/Civil-society-and-the-BWC-report-final.pdf

Lazarou, Elena, & Nicholas Lokker. 2019. *United States: Export Control Reform Act (ECRA)*. Briefing. Brussels: European Parliament Research Service. As of 27 July 2020: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/644187/EPRS_BRI(2019)644187_EN.pdf

Levine, Matt. 2015. 'Guy Trading at Home Caused the Flash Crash'. Bloomberg, 21 April. As of 27 July 2020: https://www.bloomberg.com/opinion/articles/2015-04-21/guy-trading-at-home-caused-the-flash-crash

Lewis, Dustin. 2019. 'Legal Reviews of Weapons, Means and Methods of Warfare Involving Artificial Intelligence: 16 Elements to Consider.' Humanitarian Law & Policy, 21 March. As of 27 July 2020: https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider

Microsoft. 2020. 'Responsible AI Principles from Microsoft'. As of 19 May 2020: https://www.microsoft.com/en-us/ai/responsible-ai

Monetary Authority of Singapore. 2018. *Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector*. As of 25 June 2020: https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/FEAT

NIST (US National Institute of Standards and Technology). 2020. 'Cybersecurity Framework'. As of 27 July 2020: https://www.nist.gov/cyberframework

OAS (Organization of American States) General Assembly. 2018. *Resolution on advancing hemispheric security: A multidimensional approach*, AG/RES. 2925 (XLVIII-O/18), 5 June. http://www.oas.org/en/sla/docs/AG07691E07.pdf

OECD (Organisation for Economic Co-operation and Development). 2020. 'OECD Principles on Artificial Intelligence: What are the OECD Principles on AI?' As of 19 May 2020: https://www.oecd.org/going-digital/ai/principles

OECD AI Policy Observatory. 2020. 'Countries and Initiatives Overview'. As of 3 June 2020: https://oecd.ai/countries-and-initiatives

OSCE (Organization for Security and Co-operation in Europe). 1986. *Document of the Stockholm Conference: On confidence- and security-building measures and disarmament in Europe convened in accordance with the relevant provisions of the concluding document of the Madrid meeting of the conference on security and co-operation in Europe*. As of 27 July 2020: https://www.osce.org/fsc/41238

———. 1990. *Vienna Document 1990: Of the negotiations on confidence- and security-building measures convened in accordance with the relevant provisions of the concluding document of the Vienna Meeting of the Conference on Security and Co-operation in Europe*. As of 27 July 2020: https://www.osce.org/fsc/41245

———. 2012. *OSCE Guide on Non-Military Confidence-Building Measures (CBMs)*. Vienna: OSCE. As of 19 May 2020: https://www.osce.org/secretariat/91082

———. 2013. *Permanent Council Decision No. 1106: Initial set of OSCE confidence-building measures to reduce the risks of conflict stemming from the use of information and communication technologies*, PC.DEC/1106, 3 December 2013. As of 27 July 2020: https://www.osce.org/pc/109168

Partnership on AI. 2019. *Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System*. San Francisco: Partnership on AI. As of 27 July 2020: https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system

Pecotic, Adrian. 2019. 'Whoever Predicts the Future Will Win the AI Arms Race.' *Foreign Policy*. 3 May 2019. As of 27 July 2020: https://foreignpolicy.com/2019/03/05/whoever-predicts-the-future-correctly-will-win-the-ai-arms-race-russia-china-united-states-artificial-intelligence-defense/

Permanent Mission of Switzerland to the United Nations. 2020. 'Chair's summary of the informal intersessional consultative meeting of the Open-ended Working Group on developments in the field of information and telecommunications in the context of international security', New York, 2–4 December 2019. As of 27 July 2020: https://unoda-web.s3.amazonaws.com/wp-content/uploads/2020/01/200128-OEWG-Chairs-letter-on-the-summary-report-of-the-informal-intersessional-consultative-meeting-from-2-4-December-2019.pdf

Philips. 2020. 'Philips AI Principles'. As of 19 May 2020: https://www.philips.com/a-w/about/artificial-intelligence/philips-ai-principles.html

Pontes, Giovanna F.M. n.d. 'Bridging the Gap Between Science & Diplomacy: Experiences from the OPCW'. The Hague: Organisation for Prohibition of Chemical Weapons. As of 27 July 2020: https://www.opcw.org/sites/default/files/documents/2019/07/Bridging_Science_and_Diplomacy_2019.pdf

Reiner, Philip, & Alexa Wehsener. 2019. 'The Real Value of Artificial Intelligence in Nuclear Command and Control'. War on the Rocks, 4 November 2019. As of 19 May 2020: https://warontherocks.com/2019/11/the-real-value-of-artificial-intelligence-in-nuclear-command-and-control

Ruohonen, Jukka, & Kai Kimppa. 2019. 'Updating the Wassenaar Debate Once Again: Surveillance, Intrusion Software, and Ambiguity'. *Journal of Information Technology & Politics* 16(2): 169–86. As of 27 July 2020: https://arxiv.org/pdf/1906.02235.pdf

Sarangi, Subhasish. 2019. 'National Initiatives on Artificial Intelligence in Defence'. United Service Institution of India. As of 25 June 2020: https://usiofindia.org/publication/cs3-strategic-perspectives/national-initiatives-on-artificial-intelligence-in-defence

Scharre, Paul. 2019. 'Killer Apps: The Real Dangers of an AI Arms Race'. *Foreign Affairs*, Jul/Aug. As of 19 May 2020: https://www.foreignaffairs.com/articles/2019-04-16/killer-apps

Second Review Conference of the Parties to the Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction. 1986. Final Document, BWC/CONF.II/13. https://www.unog.ch/bwcdocuments/1986-09-2RC/BWC_CONF.II_13.pdf

Shane, Scott, & Daisuke Wakabayashi. 2018. '"The Business of War": Google Employees Protest Work for the Pentagon'. *New York Times*, 4 April. As of 3 June 2020: https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html

Skitka, Linda J., Kathleen L. Mosier, Mark Burdick, and Bonnie Rosenblatt. 2000. 'Automation Bias and Errors: Are Crews Better Than Individuals?' *International Journal of Aviation Psychology* 10(1): 85–97. As of 27 July 2020: https://www.researchgate.net/publication/11803750_Automation_Bias_and_Errors_Are_Crews_Better_Than_Individuals

State Council Information Office of the People's Republic of China. 2019. *China's National Defense in the New Era*. Beijing: Foreign Languages Press Co. Ltd. As of 27 July 2020: http://english.www.gov.cn/archive/whitepaper/201907/24/content_WS5d3941ddc6d08408f502283d.html

The HCOC (The Hague Code of Conduct). 2020. 'The Hague Code of Conduct against Ballistic Missile Proliferation (HCoC)'. As of 19 May 2020: https://www.hcoc.at

Thierer, Adam. 2016. *Permissionless Innovation: The Continuing Case for Comprehensive Technological Freedom*. Fairfax: Mercatus Center, George Mason University.

Thompson, Nicholas, & Ian Bremmer. 2018. 'The AI Cold War That Threatens Us

All'. Wired, 23 October, 6.00 a.m. As of 18 May 2020: https://www.wired.com/story/ai-cold-war-china-could-doom-us-all

Tiku, Nitasha. 2018. 'The Line between Big Tech and Defence Work'. Wired, 21 May, 7.00 a.m. As of 3 June 2020: https://www.wired.com/story/the-line-between-big-tech-and-defense-work

Trapp, Ralf. 2019. *Compliance Management under the Chemical Weapons Convention*. WMDCE Series No. 3. Geneva: UNIDIR. doi: https://doi.org/10.37559/WMD/19/WMDCE3

UNGA (United Nations General Assembly). 1999a. *Developments in the field of information and telecommunications in the context of international security*, UN document A/RES/53/70, 4 January 1999. As of 27 July 2020: https://undocs.org/a/res/53/70

———. 1999b. *Review of the implementation of the recommendations and decision adopted by the General Assembly at its Tenth Special Session: Report of the Disarmament Commission*, UN document A/51/182/Rev.1, 9 June 1999. As of 27 July 2020: https://www.un.org/disarmament/wp-content/uploads/2019/09/A-51-182-Rev.1-E.pdf

———. 2015a. *Developments in the field of information and telecommunications in the context of international security*, UN document A/RES/70/237, 30 December 2015. As of 27 July 2020: https://undocs.org/A/RES/70/237

———. 2015b. *Report of the Group of Governmental Experts on Developments in the Field of information and Telecommunications in the Context of International Security,* UN document A/70/174, 22 July 2015. As of 27 July 2020: https://undocs.org/A/70/174

———. 2017. *Report of the Disarmament Commission for 2017*, UN document A/72/42, 27 April 2017. As of 27 July 2020: https://www.un.org/disarmament/wp-content/uploads/2017/06/A-72-42.pdf

———. 2018. *Developments in the field of information and telecommunications in the context of international security*, UN document A/RES/73/27, 11 December 2018. https://undocs.org/en/A/RES/73/27

UNIDIR (United Nations Institute for Disarmament Research). 2016. *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies*. Geneva: UNIDIR. As of 27 July 2020: https://www.unidir.org/files/publications/pdfs/safety-unintentional-risk-and-accidents-en-668.pdf

———. 2018a. *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies*. Geneva: UNIDIR. As of 27 July 2020: https://unidir.org/publication/algorithmic-bias-and-weaponization-increasingly-autonomous-technologies

———. 2018b. *The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence: a Primer for CCW Delegates*. Geneva: UNIDIR. As of 27 July 2020: https://www.unidir.org/files/publications/pdfs/the-weaponization-of-increasingly-autonomous-technologies-artificial-intelligence-en-700.pdf

UNODA (United Nations Office for Disarmament Affairs). 2020a. 'Developments in the field of information and telecommunications in the context of international security'. As of 19 May 2020: https://www.un.org/disarmament/ict-security/

———. 2020b. 'Military Confidence-Building'. As of 19 May 2020: https://www.un.org/disarmament/cbms

UNODC (United Nations Office for Drugs and Crime). 2011. *Model Law against the Illicit Manufacturing of and Trafficking in Firearms, Their Parts and Components and Ammunition*. Vienna: UNODC. As of 27 July 2020: https://www.unodc.org/documents/legal-tools/Model_Law_Firearms_Final.pdf

UNOG (United Nations Office at Geneva). 1993. *Ad Hoc Group of Governmental Experts to Identify and Examine Potential Verification Measures from a Scientific and Technical Standpoint*, UN document BWC/CONF.III/VEREX/9, 13–24 September 1993. As of 27 July 2020: https://www.unog.ch/bwcdocuments/1993-09-VEREX4/BWC_CONF.III_VEREX_09.pdf

———. 2020a. '2018 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS)'. As of 18 May: https://www.unog.ch/80256EE600585943/(httpPages)/7C335E71DFCB29D1C1258243003E8724

———. 2020b. 'Background on Lethal Autonomous Weapons Systems in the CCW'. As of 18 May: https://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6

US Commodity Futures Trading Commission and US Securities and Exchange Commission. 2010. *Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*. Washington, DC. As of 27 July 2020: https://www.sec.gov/news/studies/2010/marketevents-report.pdf

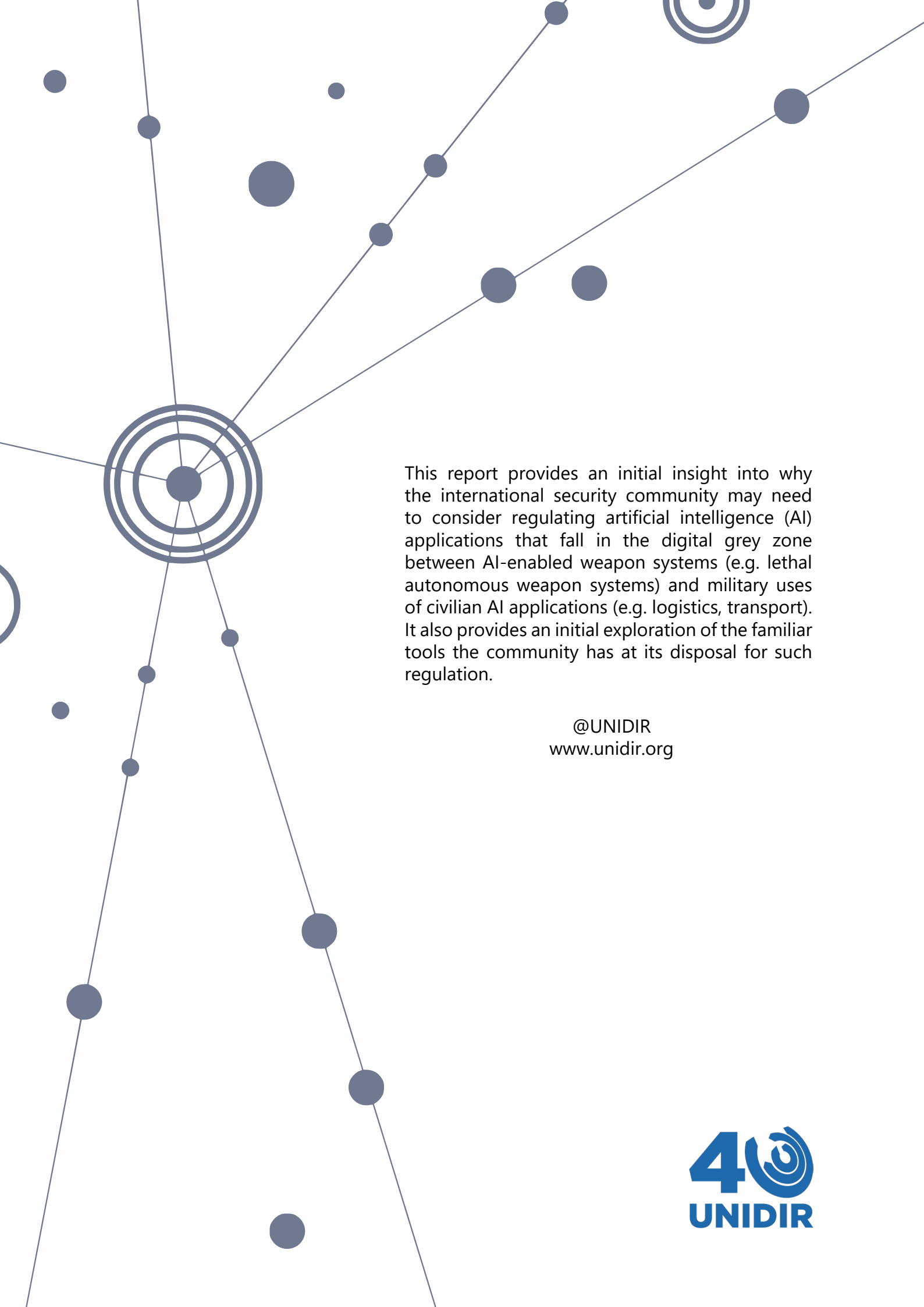
Vignard, Kerstin. 2018. 'Manifestos and Open Letters: Back to the Future?' Bulletin of the Atomic Scientists, 26 April. As of 3 June 2020: https://thebulletin.org/2018/04/manifestos-and-open-letters-back-to-the-future

Williams, Lauren. 2020. 'DOD Releases First AI Ethics Principles, but There's Work Left to Do on Implementation'. FCW, 24 February. As of 3 June 2020: https://fcw.com/articles/2020/02/24/dod-ai-policy-memo-williams.aspx

YourDictionary. 2020. 'Code of Conduct'. As of 19 May 2020: https://www.yourdictionary.com/CODE-OF-CONDUCT

Zannier, Lamberto. 2014. 'Security Community'. *Security Community* 2, 10 June: 4-6. As of 27 July 2020: https://www.osce.org/magazine/122525

Zetter, Kim. 2015. 'Why an Arms Control Pact Has Security Experts Up in Arms'. *Wired*, 24 June, 7.00 a.m. As of 28 May 2020: https://www.wired.com/2015/06/arms-control-pact-security-experts-arms



This report provides an initial insight into why the international security community may need to consider regulating artificial intelligence (AI) applications that fall in the digital grey zone between AI-enabled weapon systems (e.g. lethal autonomous weapon systems) and military uses of civilian AI applications (e.g. logistics, transport). It also provides an initial exploration of the familiar tools the community has at its disposal for such regulation.

@UNIDIR
www.unidir.org

UNIDIR