sipri

RESPONSIBLE ARTIFICIAL INTELLIGENCE RESEARCH AND INNOVATION FOR INTERNATIONAL PEACE AND SECURITY

VINCENT BOULANIN, KOLJA BROCKMANN AND LUKE RICHARDS

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE

SIPRI is an independent international institute dedicated to research into conflict, armaments, arms control and disarmament. Established in 1966, SIPRI provides data, analysis and recommendations, based on open sources, to policymakers, researchers, media and the interested public.

The Governing Board is not responsible for the views expressed in the publications of the Institute.

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE

Signalistgatan 9
SE-169 70 Solna, Sweden
Telephone: +46 8 655 97 00
Email: sipri@sipri.org
Internet: www.sipri.org

RESPONSIBLE ARTIFICIAL INTELLIGENCE RESEARCH AND INNOVATION FOR INTERNATIONAL PEACE AND SECURITY

VINCENT BOULANIN, KOLJA BROCKMANN AND LUKE RICHARDS

STOCKHOLM INTERNATIONAL
PEACE RESEARCH INSTITUTE

November 2020

Contents

Acknowledgements

This report was produced with the generous support of the German Federal Foreign Office. It is part of a research project on 'Governing the Opportunities and Risks of Artificial Intelligence for International Peace and Security', which seeks to provide input on the topic of governance of military AI in the context of Germany's efforts as part of the German presidency of the Council of the European Union and the ongoing initiative on 'Rethinking Arms Control'.*

The authors are indebted to all the experts that participated in background interviews and the participants who shared their knowledge and experience under the Chatham House Rule at the SIPRI online workshop held on 8–9 September 2020 on 'Governing the risks and opportunities of AI for international peace and security: What role for the EU?'.

The authors wish to thank the peer reviewer Charles Ovink and SIPRI colleagues Dr Sibylle Bauer, Mark Bromley, Laura Bruun, Netta Goussac and Dan Smith for their comprehensive and constructive feedback. The authors would also like to thank Moa Peldán Carlsson for her contributions in the research process that led to the production of this report. Finally, we would like to acknowledge the invaluable work of the SIPRI Editorial Department.

The views and opinions in this report are solely those of the authors and do not represent the official views of SIPRI or the funder. Responsibility for the information set out in this report lies entirely with the authors.

Vincent Boulanin, Kolja Brockmann and Luke Richards

* For information on the rethinking of arms control initiative see German Federal Foreign Office, '2020: Capturing Technology: Rethinking Arms Control', 2020. For information on Germany's presidency of the European Union see the eu2020.de website.

Executive summary

This report explores how the risks posed by the development, diffusion and military use of artificial intelligence (AI) could be mitigated through the adoption and promotion of responsible research and innovation (RRI) as an upstream approach to arms control. Its main findings and recommendations can be summarized as follows.

The development, diffusion and adoption of military and dual-use applications of AI is not inevitable; rather it is a choice, one that must be made with due mitigation of risks.

The arms control community is currently considering the role it can play in ensuring that the risks posed by AI technologies are addressed. It is still debating to what extent the standard tools of arms control can mitigate the humanitarian and strategic risks posed by the military use of AI. The fact that such use hides a complex technological reality makes the discussion on the topic challenging. AI is an enabling technology that transcends the technology-centric silos in which arms control processes usually operate. It also requires a level of technical expertise that states—as the central actors in arms control processes—might not be able to mobilize sufficiently and quickly enough to understand and react to rapid developments in this area. In addition, AI has become the object of great power competition, which adds geopolitical challenges to the pursuit of an arms control response to the risks related to military use of AI.

In this context, the report finds that RRI as an approach to technology governance could be useful for several reasons. First, it aims to involve all relevant stakeholders, particularly academia and industry, which have the technical understanding of the risks that may result from the development, diffusion and military use of AI technology. Second, it provides a governance framework for the early phase of research and development that arms control may not easily capture. Third, RRI is preventive and, by nature, iterative. It aims to identify risks and act upon them before they materialize. Moreover, it seeks to do so not just once but throughout the life cycle of technologies. Finally, because it does not necessarily aim to impose hard regulations, RRI is potentially a less politicized process than formalized arms control discussions. Like arms control, however, RRI also has its limitations. It is only one approach among others and lacks harmonized implementation and enforcement mechanisms.

At the same time, the principles and self-governance instruments that RRI creates could help the arms control community to make advances in its deliberation on the governance of the risks posed by AI. Notably, RRI processes could build on existing responsible AI initiatives, and export controls and internal compliance systems.

Many of the initiatives launched in recent years have targeted the development of principles and mechanisms for RRI in AI. These typically do not address risks related to military use of AI—although they clearly should, given the predominant dual-use nature of AI innovation. Against this backdrop, the report explores ways through which existing RRI efforts on AI could mainstream international peace and security considerations. It finds that there is a need to increase awareness about the second and third order effects of AI research and innovation, both from a humanitarian and a strategic standpoint. The report discusses how AI researchers and engineers can evaluate and limit the consequences of their work through a number of means. These could include (*a*) the implementation of very high ethical and safety standards; (*b*) the development of mechanisms and methodologies for technology impact assessment and foresight; (*c*) the design of fail-safe mechanisms; and (*d*) the application of precautionary measures in the publication of research findings. Universities, research institutes and companies already diffuse AI technology in a responsible way by complying with obligations derived from export control regulations and conducting

risk assessments required by funding organizations. Internal compliance programmes (ICPs) already provide procedures, training and systems that help researchers and developers to comply with legal provisions. In the case of AI technology, the report finds that it is a good practice to connect such compliance systems with ethical review mechanisms and robustness checks to enable a comprehensive reflection on these aspects. Ultimately, RRI should lead to decisions in the innovation and commercialization processes that can help to prevent, or pre-emptively mitigate, risks associated with the development, diffusion and military use of AI.

In the light of these findings, the report makes the following key recommendations targeted at companies, research institutes and universities that already promote or could promote RRI as a valuable approach to govern the risks posed by the military use of AI:

- Mainstream peace and security considerations into existing initiatives on responsible AI.
- Connect responsible innovation mechanisms and internal compliance programmes.

The report also makes the following recommendations aimed at states and regional organizations:

- Consider ways to consult with the AI sector in arms control discussions on AI.
- Support an initiative on responsible AI for international peace and security within the framework of the Alliance for Multilateralism.
- Identify principles for responsible military use of AI.
- Support education and training activities targeting actors in the AI sector.
- Facilitate the participation of governmental experts with military and arms control expertise in responsible AI initiatives.

Abbreviations

AI	Artificial intelligence
AI HLEG	High-level Expert Group on AI
AWS	Autonomous weapon systems
CCW	1981 Convention on Certain Conventional Weapons
DOD	Department of Defense
ELSA	Ethical, legal and social aspects
EU	European Union
FCAS	Future Combat Air System
GGE on LAWS	Group of Governmental Experts on emerging technologies in the area of lethal autonomous weapon systems
ICP	Internal compliance programme
IEEE	Institute of Electrical and Electronics Engineers
IHL	International humanitarian law
ISR	Intelligence, surveillance and reconnaissance
IT	Information technology
LAWS	Lethal autonomous weapon systems
OECD	Organisation for Economic Co-operation and Development
R&D	Research and development
RRI	Responsible research and innovation
UN	United Nations
UNODA	UN Office for Disarmament Affairs
WA	Wassenaar Arrangement
XAI	Explainable artificial intelligence

1. Introduction

Artificial intelligence (AI; see box 1) has an impact on military affairs—much in the same way as it does on people's day-to-day lives—by providing both opportunities and challenges. AI could improve the utility of future military systems by making them smarter, faster and more autonomous. At the same time, it could generate not only new humanitarian risks but also strategic ones by lowering the threshold for armed conflict, exposing civilians and civilian objects to further harm, intensifying states' (in)security and increasing the risk of crisis and conflict escalation.[1]

Over the past five years these risks have emerged as a matter of key concern for the arms control community.[2] An important topic of discussion is whether these risks can be addressed swiftly and adequately using the standard tools of arms control—and, if so, to what extent.[3] The ongoing deliberations on emerging technologies in the area of lethal autonomous weapon systems (LAWS) within the framework of the 1981 Convention on Certain Conventional Weapons (CCW) seem to indicate that attempts to develop an arms control response to military use of AI will take time and will face a number of conceptual and political challenges.[4] The CCW process on LAWS started in 2014 with the aim of determining whether autonomous weapon systems (AWS), should be specifically prohibited and regulated. Seven years later, states still disagree on the definition of AWS, the risks they pose and the policy response that is needed.[5]

As an arms control process on military use of AI might raise similar disagreements, there is a need to explore complementary approaches. In 2018 the United Nations Secretary-General identified responsible innovation of science and technology as a way to work with scientists, engineers and industry in the mitigation of risks that are posed by new technologies, and ensure their application for peaceful purposes.[6] This report explores the question of how this multi-stakeholder approach to technology governance could help to address the risks that might result from the development, diffusion and military use of AI.

The findings and recommendations are the outcome of extensive research and interviews that the authors conducted with experts in AI, responsible innovation, arms and export control, and European Union (EU) governance affairs. This report aims to inform policymakers and raise awareness within the AI community about the importance of taking into account risks related to possible military end-uses during research and innovation related to AI technologies.

Chapter 2 provides an overview of the risks posed by the military use of AI to international peace and security. It also examines whether and how these risks can be addressed by arms control. Chapter 3 introduces the concept of responsible research

[1] Boulanin, V. et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (SIPRI: Stockholm, June 2020); and Boulanin, V. et al., *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control* (SIPRI: Stockholm, June 2020).

[2] Kaspersen, A. and King, C., 'Mitigating the challenges of nuclear risk while ensuring the benefits of technology', ed. V. Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. I, *Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019).

[3] Persi Pauli, G. et al., UN Institute for Disarmament Research (UNIDIR) *Modernizing Arms Control: Exploring Responses to the Use of AI in Military Decision Making* (UNIDIR: Geneva, 2020).

[4] Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects (CCW Convention, or 'Inhumane Weapons' Convention), with Protocols I, II and III, opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983; and Brockmann, K., Bauer, S. and Boulanin, V., *Bio Plus X: Arms Control and the Convergence of Biology and Emerging Technologies* (SIPRI: Stockholm, Mar. 2019).

[5] Peldán Carlsson, M. and Boulanin, V., 'The group of governmental experts on lethal autonomous weapons systems', *SIPRI Yearbook 2020: Armaments, Disarmament and International Security* (Oxford University Press: Oxford, 2020); Rosert, E. and Sauer, F., 'How (not) to stop killer robots: A comparative analysis of humanitarian campaign strategies', *Contemporary Security Policy* (May 2020); and Kaspersen and King (note 2).

[6] UN Office for Disarmament Affairs (UNODA), *Securing Our Common Future: An Agenda for Disarmament* (UNODA: New York, 2018), pp. 52–55.

Box 1.1. What is artificial intelligence?[a]

Artificial intelligence (AI) is generally used as a catch-all term that refers to a wide set of computational techniques that allow computers and robots to solve complex, seemingly abstract problems that had previously yielded only to human cognition—for example, observing the world through vision, processing natural language and learning.[b]

From a more technical standpoint, according to the European Union (EU) High-level Expert Group on AI (AI HLEG), AI can be defined as 'software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal'.[c]

At the most basic level, AI systems always rely on the same set of technologies: sensors that collect data from the environment, a suite of computer hardware and software that allows the system to transform sensor data into purposeful plans and actions, and actuators that implement these actions on the environment. However, from one application to another, the technological components of an AI system can be fundamentally different.[d]

For the purposes of this report, AI technology could be separated into three conceptual layers: (*a*) the hardware layer, (*b*) the AI programming layer, and (*c*) the application layer.

The *hardware layer* refers to hardware components that are deemed central to the development of any type of AI system (e.g. sensors, computer chips, actuators). Most of the related technology is not AI specific. Notable exceptions are so-called AI chips that are designed specificially for AI applications such as machine learning.

The *AI programming layer* covers the design of AI technologies at the most fundamental level, before they necessarily become application specific. This corresponds to the basic AI research level where core and generic AI research problems are being explored. According to the AI HLEG, AI research problems can be grouped into three generic categories:[c]

- *Reasoning.* This category refers to techniques and research that allow humans to explicitly or implicitly programme the behaviour of the AI system. Relevant subdisciplines include knowledge representation and reasoning, planning, scheduling, search, and optimization.
- *Learning.* This category refers to machine learning techniques and approaches that allow AI systems to 'learn how to solve problems that cannot be precisely specified, or whose solution methods cannot be described by symbolic reasoning rules'.[c] Relevant subdisciplines include neural networks, deep learning, supervised machine learning, unsupervised machine learning and reinforcement learning.
- *Robotics or embodied AI.* This category refers to research problems aimed at allowing AI systems to act in the physical world. Problems in the reasoning and learning categories (mentioned above) are relevant to robotics. Robotics is, however, distinct from AI because it also mobilizes other disciplines and research problems in the areas of mechanical engineering and control theory.

The application layer is where AI systems find a concrete application and commercial end-use. It can be further subdivided into two interchangeable categories:

- *By sector*: such as health, education, business administration, industrial automation and military; or
- *By type of task or application*: such as computer vision, natural language processing, problem solving, data management systems (classification, optimization, prediction, anomaly detection etc.) and robotics.

[a] For definitions of some of the key terms used here see e.g. Russell, S. and Norvig, P., *Artificial Intelligence: A Modern Approach*, 3rd edn (Pearson Education: Harlow, 2014); and Montreal Ethics Institute, 'AI Ethics Living Dictionary', [n.d.].

[b] International Panel on the Regulation of Autonomous Weapons (IPRAW), *Focus on Computational Methods in the Context of LAWS*, 'Focus on' Report no. 2 (German Institute for International and Security Affairs: Berlin, Nov. 2017).

[c] For the full definition and the reasoning behind it see High-level Expert Group on AI (AI HLEG), *A Definition of AI: Main Capabilities and Disciplines* (European Commission: Brussels, 2019), p. 3.

[d] Russell and Norvig (note a).

and innovation in the context of AI, and explains how and why it could help to achieve arms control objectives on the military use of AI. Chapter 4 maps existing efforts with regard to responsible development, diffusion and use of AI technology, through 'responsible AI initiatives', export controls and compliance systems. It discusses the main challenges for responsible research and innovation in the context of AI and highlights opportunities that could be built on. The concluding chapter (chapter 5) summarizes the key findings of the report and outlines specific recommendations for research institutes and universities as well as for industry, states and the EU.

2. Addressing the risks posed by the military use of AI

I. AI and international peace and security

Over the past decade, AI has achieved notable technological breakthroughs, largely thanks to improvements in a technique called machine learning.[7] In the military realm, recent advances in AI could strengthen the capabilities of armed forces across the board—from intelligence, surveillance and reconnaissance (ISR) through to combat operations and logistics—while allowing for more cost efficiency (see figure 2.1).[8] There is a rapidly expanding body of literature that addresses the impact that the increasing use of AI in military systems could have on international peace and security.[9] This discussion is still at a relatively early stage and is generally focused on two categories of risk—humanitarian and strategic. These risks can themselves result from the way AI technology is developed, diffused or used.

Humanitarian and strategic risks

The risk from a humanitarian perspective is that AI could, by design or through the way it is employed, undermine the ability of the military to operate within the limits of international humanitarian law (IHL). This, in turn, could expose civilians and civilian objects to greater risk of harm, death or destruction.[10] This concern is already central to the debate on emerging technologies in the field of LAWS at the CCW. CCW states parties have discussed whether the use of AI to increase autonomy in weapon systems could diminish a military commander's ability to foresee the consequences of the use of force in an attack. This could undermine the commander's ability to properly exercise the context-specific evaluative judgements that IHL demands and could potentially lead to violations of IHL. However, the military use of AI is not limited only to autonomy in weapon systems; experts are also concerned that the use of AI in decision support systems could be problematic if adopted without proper safeguards in place. Known design flaws such as data bias and algorithmic opacity could induce users to make mistakes or misjudge situations—which could have severe humanitarian consequences.[11]

From a strategic standpoint, experts are concerned that AI may have destabilizing effects on international peace and security. The increasing military use of AI could undermine states' sense of security, while introducing new variables of non-

[7] Machine learning is a collective name often used for statistical methods of identifying structures in data. For more detail see Hagström, M., 'Military applications of machine learning and autonomous systems', ed. Boulanin (note 2).

[8] Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017); Boulanin et al., *Artificial Intelligence, Strategic Ability and Nuclear Risk* (note 1); and Scharre, P. and Horowitz, M. C., 'Artificial intelligence: What every policymaker needs to know', Center for New American Security, 19 June 2018.

[9] Horowitz, M. C. et al., 'Strategic competition in an era of artificial intelligence', Center for New American Security, 25 July 2018; Cummings, M. L., *Artificial Intelligence and the Future of Warfare* (Chatham House: London, Jan. 2017); and Roff, H. M. and Moyes, R., 'Meaningful human control, artificial intelligence and autonomous weapons', Briefing Paper prepared for Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons, Article36.org, 8 Apr. 2016.

[10] Schmitt, M. N. and Thurnher, J. S., '"Out of the loop": Autonomous weapon systems and the law of armed conflict', *Harvard National Security Journal*, vol. 4, no. 2 (May 2013); International Committee of the Red Cross (ICRC), *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?* (ICRC: Geneva, Apr. 2018); and Roff and Moyes (note 9).

[11] Schmitt and Thurnher (note 10); ICRC (note 10); Roff and Moyes (note 9); Boulanin et al., *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control* (note 1); Asaro, P., 'On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making', *International Review of the Red Cross*, vol. 94, no. 886 (summer 2012); and United Nations, General Assembly, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns, A/HRC/23/47, 9 Apr. 2013.

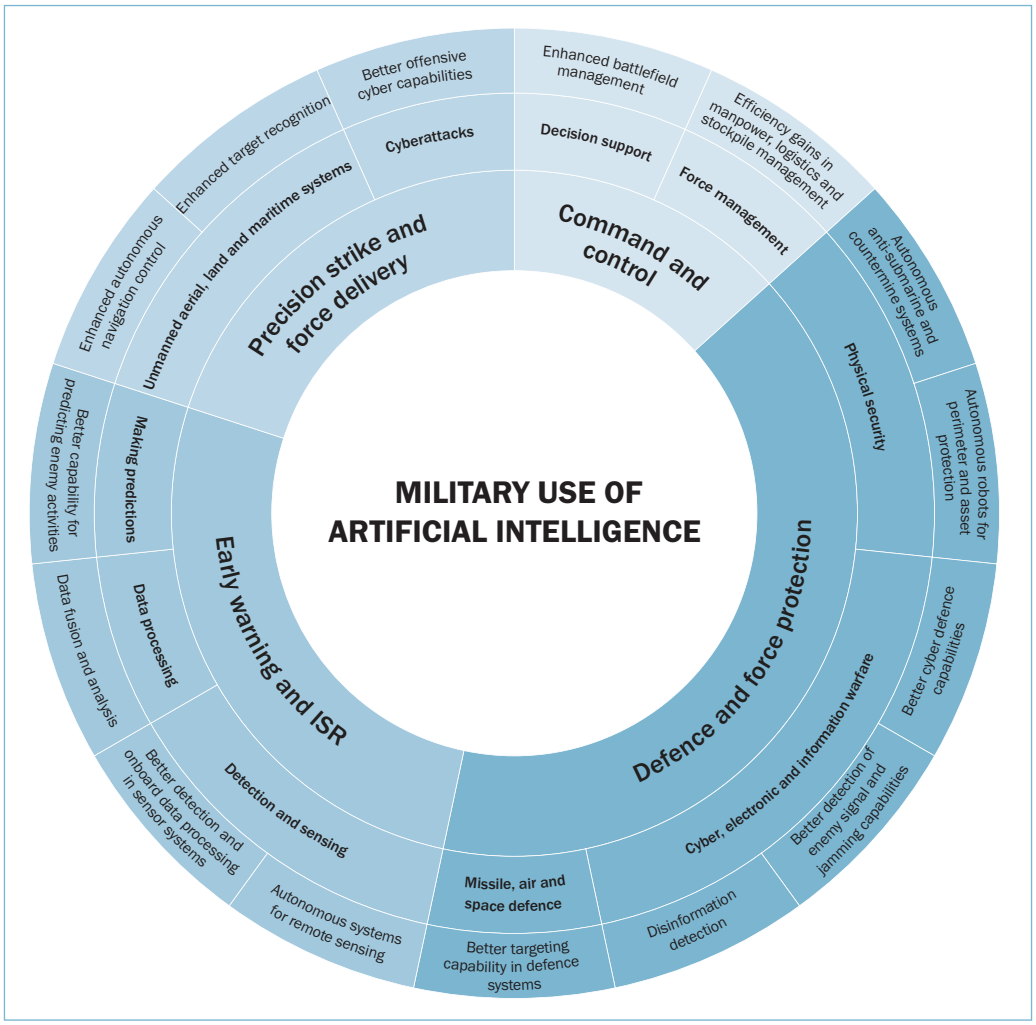

Figure 2.1. Foreseeable military applications of AI

ISR = intelligence, surveillance and reconnaissance.

Source: Boulanin, V. et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (SIPRI: Stockholm, 2020).

transparency and unpredictability into the strategic relations between states. A race to develop AI technologies with military applications could also widen the digital gap (i.e. the access to and possibility to use digital technologies) between certain states and, in general terms, could lead to the diversion of more resources to the production of arms. Experts are also concerned that the increasing military use of AI could lower thresholds for violence, reduce opportunities for de-escalation and raise the risk of proliferation to unauthorized end-users or to actors who may use the technology in a destabilizing way.[12]

[12] Altmann, J. and Sauer, F., 'Autonomous weapon systems and strategic stability', *Survival*, vol. 59, no. 5 (Nov. 2017); Horowitz, M. C., 'Artificial intelligence, international competition, and the balance of power', *Texas National Security Review*, vol. 1, no. 3 (May 2018); Gates, J., 'Is the SSBN deterrent vulnerable to autonomous drones?', *RUSI Journal*, vol. 161, no. 6 (2016); and Hambling, D., 'The inescapable net: Unmanned systems in anti-submarine warfare', *Parliamentary Briefings on Trident Renewal*, no. 1, British–American Security Information Council (BASIC), 13 July 2016; Geist, E. and Lohn, A. J., *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (RAND Corporation: Santa Monica, CA, 2018); Boulanin et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (note 1); and Rickli, J., 'The impact of autonomy and artificial intelligence on strategic stability', *UN Special*, no. 781 (July–Aug. 2018).

Box 2.1. AI explainability and the black box problem

Machine learning algorithms can often operate like a black box—that is, while the input and the output of such a system are observable, the computational process undertaken cannot be fully explained. This also makes it difficult to know what a system has learned and how it would react to data outside that on which it has been trained.[a] Thus, there is a need to ensure that such systems are less opaque so as to align them with human values. Moreover, potential biases within artificial intelligence (AI) training data can affect the decision making of these systems, compounding the need for transparency and explainability of AI systems.[b]

Technical explainability may require a decision made by an AI system to be understandable, leading to a possible trade-off between an algorithm's accuracy and its explainability.[c] However, examples of interpretable models already exist and in some cases interpretability constraints can potentially be added without reducing a model's accuracy.[d] Efforts are being made in explainable AI (XAI) research programmes to be able to extract the relevant elements the various stakeholders need to understand the way in which an AI system works.[e]

[a] Righetti, L., 'Emerging technology and future autonomous systems', International Committee of the Red Cross (ICRC), *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, Expert meeting, Versoix, Switzerland, 15–16 Mar. 2016 (ICRC: Geneva, Aug. 2016), pp. 36–39.

[b] Pedreschi, D. et al., 'Meaningful explanations of black box AI decision systems', *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1 (2019).

[c] High-level Expert Group on Artificial Intelligence (AI HLEG), *Ethics Guidelines for Trustworthy AI* (European Commission: Brussels, Apr. 2019), p. 18.

[d] Rudin, C. and Radin, J., 'Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition', *Harvard Data Science Review*, vol. 1, no. 2 (Nov. 2019).

[e] Zednik, C., 'Solving the black box problem: A normative framework for explainable artificial intelligence', *Philosophy and Technology* (2019); and Turek, M., 'Explainable artificial intelligence', Defense Advanced Research Projects Agency, [n.d.].

Risk vectors: Development, diffusion and use of AI technology

The humanitarian and strategic risks outlined above may derive from (*a*) the way AI technology is developed (design-induced risks), (*b*) diffused (diffusion-induced risks), (*c*) used (risk of misuse), or (*d*) a combination of these factors.

Design-induced risks are those resulting from choices or non-choices made in the development process. Choices that affect the transparency, understandability, explainability and reliability of AI systems can determine the extent to which such systems may be used in accordance with IHL (see box 2.1). They also have the potential to spark incidents that could be destabilizing for strategic relations between states and that could possibly trigger conflicts.[13] One scenario that has been discussed is the prospect of a modern version of the 1983 Petrov incident, whereby an early warning system powered by machine learning would wrongly identify that an attack is under way and force the military command to decide within minutes whether to respond to that attack.[14] The opacity of machine learning algorithms could not only cause reliability problems to stay undetected, but also complicate the commander's responsibility for verifying the information provided by the system.

Diffusion-induced risks result from the extent to which the technology, or knowledge to design it, is available to states and other actors pursuing military or other potentially destabilizing end-uses. The proliferation of AI military technology could in some regions undermine relations between states. Another challenge is proliferation to unauthorized and irresponsible actors who could use the technology to threaten populations or the security of a state.

Finally, the risk of misuse is inherent to any military technology. Any military or dual-use technology may be used in a way that creates humanitarian and strategic concerns. For example, from a humanitarian standpoint, a state or non-state actor could misuse

[13] Boulanin et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (note 1).

[14] The 1983 Petrov incident occurred on 26 Sep. 1983. The Soviet Union's early warning system wrongly identified that a US missile attack was under way. Lieutenant Colonel Stanislav Petrov eventually decided not to report the incident as he believed it was a false alarm. It turned out that a satellite had wrongly identified the missiles. Petrov's decision was later hailed as having saved the world from nuclear war. Boulanin et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (note 1), pp. 21, 114–16.

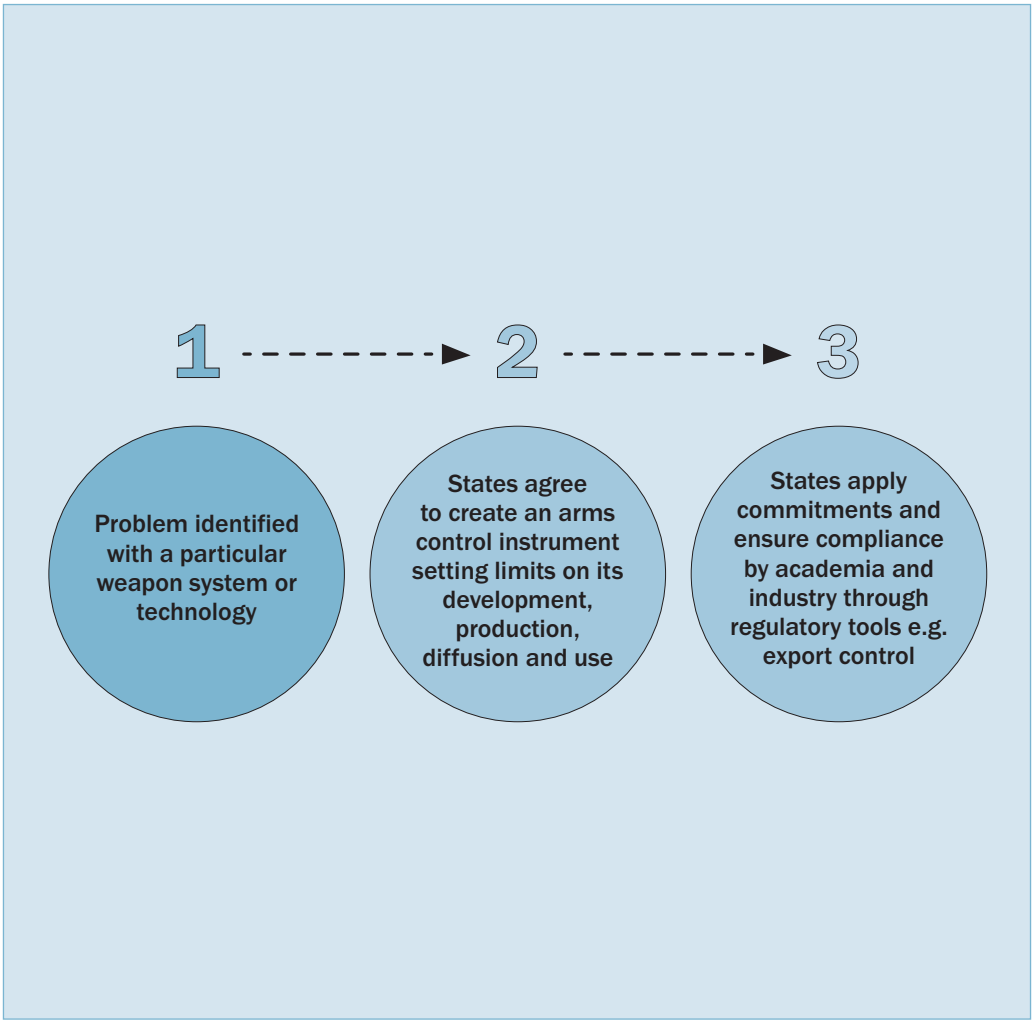

Figure 2.2. Arms control as a process

Source: Authors' conceptualization.

AI technology by employing AI-powered AWS with limited distinction capabilities and no human oversight, and accidentally target civilians and civilian objects. From a strategic standpoint, a state or non-state actor could misuse AI technology to generate deepfakes to spread disinformation and destabilize the political systems of a country or fuel tension between two countries.[15]

II. Addressing humanitarian and strategic risks using arms control

The development, diffusion and adoption of military and dual-use applications of AI is not inevitable; rather it is a choice, one that must be made with due mitigation of risks. The humanitarian and strategic risks posed by the military use of AI can and need to be addressed—the question is how this can be achieved. This question has become a matter of concern for the community of experts who work with international security issues and arms control. It is generally agreed that the development and military use of AI is not taking place in a governance vacuum. International law, particularly IHL, sets clear limits on what may be deemed responsible development and use of military technology. Arms control could provide tools for further governing the development,

[15] Fitzpatrick, M., 'Artificial intelligence and nuclear command and control', Survival Editor's Blog, International Institute for Strategic Studies (IISS), 26 Apr. 2019.

use and diffusion of military AI. However, arms control processes have limitations. These are the focus of this section.

Arms control as a tool to govern the development, diffusion and military use of AI

Assessing the need to develop an arms control response to the military use of AI

International law governs the military use of AI in armed conflicts. Article 36 of 1977 Additional Protocol to the 1949 Geneva Conventions (Additional Protocol I) obligates states to ensure that, in the study, development, acquisition and adoption of any new weapon or means or method of warfare, international law does not prohibit or restrict its employment.[16] This obligation applies to states that develop weapons as well as those that import weapons from other countries.

International law, particularly IHL, has determined how weapons, and means and methods of warfare—regardless of their type—may or may not be used. However, the debate on emerging technologies in LAWS at the CCW has shed light on the fact that AI might pose novel and unique challenges. IHL, for instance, does not explicitly require human control over the use of force, while the use of AI to increase autonomy in weapon systems could potentially prevent a commander from foreseeing and limiting the consequences of the use of force in an attack.[17] In that context, the debate is unresolved on whether there is a need to clarify or develop IHL by introducing new humanitarian arms control tools. Some states and civil society groups see a need for new regulation to clarify or develop existing rules of IHL, while others believe IHL is sufficient as it already prescribes a complex set of limits.[18]

At the same time, states and civil society groups recognize that IHL and humanitarian arms control cannot cover the entire spectrum of issues posed by the military use of AI. Strategic risks might require a dedicated response in the form of bilateral or multilateral limits on specific AI-related capabilities.

In summary, the arms control community is already considering ways in which arms control can be used to ensure that the risks posed by the military use of AI technologies are addressed.[19] Although the debate on this topic is still nascent, there already seems to be a consensus that using arms control to govern the development, proliferation and use of military AI will not be easy to implement. There are three major challenges—of a conceptual, sequencing and political nature—to overcome.[20]

Conceptual challenge: Defining military AI and the problem it poses

Typically, arms control processes are ex-post processes—that is, they are developed in reaction to actual events or at least to a well-identified problem. States agree on these controls and then incentivize or enforce regulatory means on relevant actors (from research, academia and industry) for implementation (see figure 2.2).

[16] Boulanin, V. and Verbruggen, M., *Article 36 Reviews: Dealing With the Challenges Posed by Emerging Technologies* (SIPRI: Stockholm, Dec. 2017).

[17] Boulanin et al., *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control* (note 1).

[18] Peldán Carlsson and Boulanin (note 5).

[19] Arms control traditionally refers to mutually agreed upon restraints or controls (usually between states) on 'development, production, stockpiling, proliferation, deployment and use' of states' 'military capability or potential'. Controls can be bilateral (between two states) or multilateral (between several states). Examples of multilateral arms control include the prohibitions on biological and chemical weapons, anti-personnel landmines and cluster munitions. States operationalize these commitments by placing limits on the development and deployment of particular weapon systems, disposing of any stockpiles, and making adjustments to the scope and implementation of their export controls. For further detail see e.g. North Atlantic Treaty Organization (NATO), 'Arms control, disarmament and non-proliferation in NATO', 16 Mar. 2020.

[20] Kaspersen and King (note 2).

The challenge with this model is that, as a baseline, states need to have a common understanding—both internally and between each other—of the essence and extent of the problem in order to create an arms control instrument. This is difficult in the case of AI because the technology is intangible, multipurpose in nature and complex.

To date, there have been no actual events or consequences that could serve as a baseline for defining a problem and building consensus around it, unlike the case of the prohibition of biological and chemical weapons or anti-personnel landmines and cluster munitions. Currently, there are no weapon systems in active service that have the capacity to deliver lethal force and use AI powered by machine learning. One of the main reasons for this is that machine learning algorithms are opaque and therefore difficult to certify for safe use (see box 2.1).[21]

The arms control community has demonstrated in the past that it can be forward looking and can take action before a weapon or capability is developed and used. One example of this type of preventive arms control is the CCW protocol on blinding laser weapons.[22] However, in the case of military AI, it is hard to formulate one clearly identifiable overarching problem. AI is an enabling technology that has not one but many possible military uses, of which only some may generate the aforementioned humanitarian and strategic risks. The arms control community would need to consider the risks posed by AI—and hence the governance response—in relation to specific military-related applications.[23]

Furthermore, 'military use of AI' as an abstract term hides a complex reality, which can be difficult to communicate in multidisciplinary settings and multilateral diplomatic negotiations.[24] It naturally takes time for states, especially those that might not have the relevant technical expertise readily available, to understand and assess the technology and its implications at a more granular level. The technical complexity and the fact that states might have different levels of understanding of the technology are major obstacles to consensus building.

The multipurpose nature of AI technology is a source of concern as arms control issues have traditionally only been discussed and addressed in institutional silos, such as within the framework of specific UN conventions or UN bodies such as the Conference on Disarmament, which are limited by their mandate in terms of topic and process.[25] All the areas covered by arms control—conventional, nuclear, chemical and biological or cyber weapons and related capabilities—in theory could leverage technological advances in the field of AI. The question of whether the relevant institutional silo should deal with the challenges posed by AI in each specific area or whether a separate, dedicated process is needed remains the subject of debate. However, the conceptual reasons outlined above indicate that the creation of an overarching arms control process dedicated to the whole range of AI applications would be difficult to realize and at this point it appears highly unlikely that such a process could be implemented.

Sequencing challenge: Keeping up with the pace of advances in AI

Many parts of the AI field are progressing rapidly as a result of increased algorithmic efficiency along with growing computational power and data availability, widening

[21] Hagström (note 7); and Boulanin and Verbruggen (note 8).

[22] Rosert and Sauer (note 5).

[23] In concrete terms, this means that a general arms control discussion on military AI might be impractical and that discussions might need to take place in multiple forums depending on areas of application: conventional, cyber, and nuclear.

[24] Boulanin, V., *Mapping the Debate on LAWS at the CCW: Taking Stock and Moving Forward*, EU Non-proliferation Paper no. 49 (SIPRI: Stockholm, Mar. 2016).

[25] Brockmann, Bauer and Boulanin (note 4); and Kaspersen and King (note 2).

the scope of the training of algorithms.[26] These advances are being driven by significant private sector investment and a huge military appetite for the technology. In contrast, arms control processes and negotiations usually move slowly. With the notable exception of the CCW protocol on blinding laser weapons, it typically takes many years—and in numerous instances decades—for arms control processes to result in concrete outcomes. Two significant examples of relevant processes that touched upon the issue of AI are the UN processes on the development of international norms of responsible state behaviour in cyberspace (since 1998) and on LAWS (since 2014).[27] Progress in these processes has been exceedingly slow, both at the substantial level (e.g. determining what the problem is and the interpretation of the applicability of IHL) and at the political level (e.g. agreeing on the desired political outcomes). Therefore, there are legitimate concerns that advances in AI could outpace any arms control process.[28] If policymakers lag behind technological developments, there is a risk that new applications may be adopted without appropriate safeguards in place. Some technologies and their use might also be difficult to govern once they are adopted and used by some militaries.

Political challenge: Finding agreement between states

Arms control processes are also highly state centric. States are the one entity that have the power to determine where and when regulations are needed; what rules should apply, to whom and how. The prospects for arms control on AI are therefore contingent on the political will of governments. Finding agreement among states on AI governance will probably be difficult in the current geopolitical context. [29] In recent years states have increasingly questioned the role of arms control as a mechanism for promoting peace and security. Major powers, including China, Russia and the United States, currently appear to have limited faith in each other's engagement in arms control processes.[30] These states also have a vested interest in not limiting the speed and trajectory of developments in AI technology. They are therefore likely to object to any initiative that could cause them to lose a perceived advantage, or become somehow disadvantaged, in their strategic competition. The current 'arms control winter' in combination with the great power competition on AI, makes the chances of an arms control agreement on military use of AI remote—at least for the time being.[31]

The need to develop complementary responses to the military use of AI

In summary, the military use of AI poses a number of humanitarian and strategic concerns that demand a response focused on the risks that might emerge from the development, diffusion and military use of AI technology. There are, however, some challenges associated with using arms control processes to address these risks. These

[26] Amodei, D. and Hernandez, D., 'AI and compute', OpenAI, 16 May 2018; and Hernandez, D. and Brown, T., 'AI and efficiency', OpenAI, 5 May 2020.

[27] For an overview of the UN process on the development of international norms of responsible state behaviour in cyberspace see e.g. Davis, I. et al., 'Conventional arms control and new weapon technologies', *SIPRI Yearbook 2020* (note 5). The process led to two parallel initiatives in 2019: the Open-ended Working Group and a new Group of Governmental Experts.

[28] See e.g. Kaspersen and King (note 2).

[29] Kühn, U., 'Why arms control is (almost) dead', Strategic Europe, Carnegie Europe, 5 Mar. 2020; and Pifer, S., 'As US–Russian arms control faces expiration, sides face tough choices', Order From Chaos Blog, Brookings, 23 Mar. 2020.

[30] Countryman, T., 'Why nuclear arms control matters today', *Foreign Service Journal* (May 2020); Asada, A., 'A "winter phase" for arms control and disarmament and the role for Japan', *Japan Review*, vol. 3, no. 3–4 (spring 2020); and Sauer, F., 'Stepping back from the brink: Why multilateral regulation of autonomy in weapons systems is difficult, yet imperative and feasible', *International Review of the Red Cross*, Special Issue on Digital Technologies and War (forthcoming).

[31] The term 'arms control winter' was coined by Frank Sauer to refer to the political deadlock in which arms control finds itself as a result of the (dis)engagement of the major powers. The idea is that, in the current political climate, there is little prospect for new arms control ideas to grow. See Sauer (note 30); and Borrie, J., 'Cold war lessons for automation in nuclear weapon systems', ed. Boulanin (note 2).

challenges do not mean that arms control on military AI is irrelevant or would be unable to provide an effective governance mechanism; rather they show that other complementary processes might need to be explored. These are the focus of the next chapter.

3. Responsible research and innovation as a means to govern the development, diffusion and use of AI technology

Given the leadership of the civilian sector in AI innovation and the state-centric nature of multilateral arms control, multi-stakeholder initiatives—involving representatives from research, academia, the private sector, government and civil society—could be useful to help to address the risks to international peace and security posed by the development, diffusion and military use of AI. One option could be to follow the recommendation of the UN Secretary-General who identified responsible innovation as a way to work with researchers, academia and the private sector on the mitigation of risks posed by emerging technologies.[32] This chapter outlines why and how responsible research and innovation (RRI), as an approach to technology governance, is valuable for pursuing arms control objectives on the military use of AI.

I. RRI in the support of arms control on the military use of AI

RRI as an approach to technology governance

RRI is a relatively new concept that has been defined as a:

> transparent, interactive process by which societal actors and innovators become mutually responsible to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society).[33]

In concrete terms, RRI aims to allow:

> all stakeholders that are involved in the processes of research and innovation at an early stage (A) to obtain relevant knowledge on the consequences of the outcomes of their actions and on the range of options open to them … (B) to effectively evaluate both outcomes and options in terms of societal needs and moral values and (C) to use these considerations (under A and B) as functional requirements for design and development of new research, products and services.[34]

RRI emerged as a replacement to the ethical, legal and social aspects (ELSA) of research and innovation within EU policy discourse and practice.[35] This change caused a shift towards an approach to research and innovation that is anticipatory, in the absence of certitude around both the impact of emerging science and technologies and the coverage of existing regulations.[36] It is one way to concretely enact responsible innovation in science and technology, the need for which was espoused by the UN Secretary-General. Some of the earliest incarnations of an RRI agenda can be found in the field of nanotechnology.[37]

[32] UN Office for Disarmament Affairs (note 6).

[33] Von Schomberg, R., 'A vision of responsible research and innovation', eds R. Owen, M. Heintz and J. R. Bessant, *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society* (Wiley: London, 2013).

[34] Van Oudheusden, M., 'Where are the politics in responsible innovation? European governance, technology assessments, and beyond', *Journal of Responsible Innovation*, vol. 1, no. 1 (2014), p. 70. See also European Commission, Directorate-General for Research and Innovation, *Options for Strengthening Responsible Research and Innovation*, Report of the Expert Group on the State of Art in Europe on Responsible Research and Innovation (European Commission: Brussels, 2013), Annex I, p. 56.

[35] Zwart, H., Landeweerd, L. and van Rooij, A., 'Adapt or perish? Assessing the recent shift in the European research funding arena from "ELSA" to "RRI"', *Life Sciences, Society and Policy*, vol. 10, no. 11 (2014).

[36] Van Oudheusden (note 34).

[37] The operationalization of RRI within this context can be seen in a number of case studies, such as the NanoSoc project, which has the aim of increasing new forms of collaboration to drive innovation and R&D with broad public support, and Applied Nanoparticles SL, a company established for researching, studying and developing nanoparticles and their applications based on RRI principles. For further detail see Busquets-Fité, M. et al., *RRI Case Study: Applied*

In practical terms, RRI as a technology governance approach can be implemented through different instruments, which can be established within or between universities, research institutions and companies. These instruments include the following:

1. Ad hoc or permanent groups where relevant stakeholders from various organizations or disciplines (e.g. social science and natural science) can discuss or monitor desirable and undesirable outcomes of innovation. Examples include ethical review boards within research institutes, universities, companies and governments; research and development (R&D) funding organizations; and project selection committees connected to R&D funding programmes.
2. Guidelines and principles that set the baseline of research and innovation outcomes that could be problematic from a legal, ethical and safety standpoint.
3. Codes of conduct that promote responsible behaviour for relevant stakeholders in research, academia, the private sector and governments.
4. Industry standards that define baseline levels of safety required for the development and testing of new technologies that are iterative and responsive to change.
5. Methodologies for technology impact assessments and forecasting, which might come from, among other things, expert elicitation (i.e. interdisciplinary expert discussion), scenario planning or formal modelling.
6. Capacity-building and training mechanisms on RRI, ethics and the societal impact of science and technology targeted at academia, companies or governments at the macro level, or individuals or groups of researchers and engineers at the more micro level.[38]

The advantages of RRI for technology governance and arms control

RRI as an approach to technology governance is valuable for the pursuit of arms control objectives for several reasons, which can be grouped into three broad categories. RRI is (*a*) comprehensive, inclusive and technology specific, (*b*) reflexive and preventive, and (*c*) principles based. These features, which are discussed in more detail below, mean that RRI has the potential to address or bypass the three key challenges identified in chapter 2—conceptual, sequencing and political—faced by the arms control community in the governance of military uses of AI.

A comprehensive, inclusive and technology-specific approach

RRI is a 'comprehensive approach of proceeding in research and innovation' that aims to identify problems throughout the life cycle of technology: from basic scientific research to product commercialization.[39] In addition, RRI is iterative, as it seeks to address and understand potential problems over the life cycle of science and technology. From this perspective, RRI is useful to address part of the sequencing challenge discussed in chapter 2. It could provide a framework for governing technological innovation that may not be explicitly regulated by arms control, from the early phase of R&D to commercialization.

RRI is also an inclusive and multi-stakeholder approach. It aims to involve a diverse array of actors in research, academia, the private sector and government. From an arms control perspective, such inclusiveness is valuable because it could create or improve

Nanoparticles SL: Spinning Off under Responsible Research and Innovation (RRI) Principles (Responsible Innovation Compass: 2020).

[38] Brundage, M., 'Responsible governance of AI: An assessment, theoretical framework, and exploration', PhD dissertation, Arizona State University, Dec. 2019.

[39] Van Oudheusden (note 34), p. 70.

the link between the technology innovators and the arms control community—which is, as previously discussed, highly state centric. Arms control discussions on LAWS, for example, have demonstrated that it can be difficult for the arms control community to fully leverage technical expertise from the AI sector in the debate. At the same time, as discussed in the next chapter (chapter 4), the civilian actors in the AI sector seem to have limited awareness of the possible impact of AI innovation in the military sector. RRI's inclusiveness provides an essential opportunity to connect relevant communities, which could be particularly useful in the light of the conceptual challenges discussed in chapter 2. The interaction between a diverse community is essential to ensuring that the risks associated with the military use of AI are accurately identified—in a way that neither underestimates nor overestimates them and does not overlook issues that only certain types of actors might notice. This interaction is also essential to ensuring that risk management responses are appropriately selected and that decisions to limit, or not to pursue, specific research and innovation processes do not create new or excessive negative political, economic or societal outcomes for some of the actors involved (e.g. create unfair disadvantages between countries, companies or societies).

RRI is also technology specific. The risks are identified and addressed in relation to specific technologies or areas of application such as, for example, facial recognition or autonomous systems. This could be valuable in addressing the conceptual difficulty that AI poses to the arms control community. The enabling and multipurpose nature of AI technology makes top-down governance approaches difficult to develop and implement. RRI helpfully provides a bottom-up approach to risk identification. It grounds the discussion in specific developments of science and technology.

A reflexive and preventive approach

RRI is also a reflexive and preventive approach, which makes it useful to tackle the sequencing challenge that arms control typically faces with regard to the governance of new and rapidly advancing technologies. RRI invites relevant actors to identify and respond to problems before they occur—not only through design choices but also through self-restraint in the diffusion and trade of the products of research and innovation. RRI provides an opportunity for identifying and addressing risks before they materialize. Another related valuable feature is that RRI is grounded in ethics and social desirability rather than hard law. While RRI processes take existing national and international regulations as a baseline, they also provide means for researchers and engineers to make decisions in the absence of clear regulations.[40]

RRI could perhaps be described as an upstream aspect of arms control because the processes it creates might allow for (*a*) early interventions on technological developments that might create arms control concerns, (*b*) early identification of issues that might require political or regulatory responses from the arms control community, and (*c*) identification of principles and good practices for the development, diffusion or use of technology that could feed into traditional arms control processes aimed at regulating the military use of AI.

It is worth stressing that RRI will not be a silver bullet for all the risks that AI poses to international peace and security. It provides a useful set of tools and processes but cannot by itself effectively address all types of risk. Some issues will necessarily require governments to provide political responses at the national or international level. RRI processes could, nonetheless, work as indicators for identifying problems that may deserve dedicated multilateral arms control processes or a regulatory response.

[40] Van Oudheusden (note 34).

A principles-based approach

RRI as an approach to technology governance is valuable as it could help to bypass the political deadlock in which multilateral arms control currently finds itself. The political strength of RRI is that it aims to provide non-legally binding principles and best practices. As such, it is not politicized in the same way as arms control deliberations. There is no threat of regulation, which might make actors fear for their own interests and refuse to engage constructively in the discussion. This is not to say that RRI processes will not be politicized. RRI processes necessarily involve an interaction between different value judgements and interests, which has the potential to cause conflict. However, because RRI processes do not explicitly aim to create hard regulations, the political stakes in them for the various actors, might not be as high.

A related advantage of RRI is that it does not simply aim to limit the negative effects of innovation using a form of 'precautionary principle'; rather, it is also intended to enable a reflection on what constitutes a positive effect of technology. In other words, it is geared towards generating positive outcomes, not just preventing negative ones.[41] The 'positive' framing of RRI could be more acceptable than, for instance, preventive arms control to actors that might have political reservations against processes that could limit their ability to develop new technologies and leverage them for military purposes. As such, RRI is valuable as it could provide an alternative framework for engaging actors that are wary of arms control processes (not only states but also civilian actors that might not be willing to discuss arms control issues; see chapter 4) in the pursuit of arms control objectives.

II. How would RRI in AI work in practice?

The question then is how RRI in AI, with a view to mitigating the risks for international peace and security, would work in practice. This section suggests some pathways through which the consideration of risks for international peace and security could be integrated into existing and future RRI processes on AI. Based on the practical description of the operation of RRI noted at the beginning of this chapter, this section outlines (*a*) the knowledge that AI actors might need to evaluate the outcome of their work in the context of peace and security, (*b*) the means that could be used to support them in their evaluation, and (*c*) some of the possible outcomes of an RRI process.[42]

The knowledge needed

Defining responsible AI in the context of international peace and security

A necessary first step towards identifying the knowledge that AI actors might need to evaluate the outcome of their work is to define what 'responsible' innovation in AI means in the context of international peace and security. Finding a narrative that makes a concrete connection between the development and diffusion of AI technology on the one hand and peace and security on the other can be difficult given the variety of technologies and complexity of the risk scenarios involved. It is possible, however, to depict at the more fundamental level what the pillars of responsible innovation should

[41] Arnaldi, S., Gorgoni, G. and Pariotti, E., 'RRI as a governance paradigm: What is new?', eds R. Lindner et al., *Navigating Towards Shared Responsibility in Research and Innovation Approach: Process and Results of the Res-AGorA Project* (Fraunhofer Institute for Systems and Innovation Research ISI: Karlsruhe, 2016), p. 27. The precautionary principle is often invoked where there is uncertainty with regard to a phenomenon, product or process and the consequences are not fully understood, and there is a potential need to take pre-emptive measures. For an overview of some of its uses see the various articles on this topic published by *ScienceDirect*. The principle also has a specific context within EU law. For further detail see European Commission, 'Communication from the Commission on the precautionary principle', COM(2000) 1 final, 2 Dec. 2000.

[42] Van Oudheusden (note 34), p. 70.

be. In 2019 the High-level Expert Group on AI (AI HLEG), which had been convened by the European Commission, put forward recommendations on trustworthy AI that were premised on three components—namely that it should be legal, ethical and robust (i.e. reliable and safe from a technical standpoint).[43] These three components provide a conceptual framework and pillars for considering the elements of responsible innovation as they relate to international peace and security.

First, AI technology should be developed and diffused with regard for the limits and requirements that exist in international law—IHL and human rights law—in addition to any national legal obligations, in accordance with commitments made in arms control and export control forums.[44]

Second, AI technology should be developed while keeping in mind ethical considerations. The issue here is that, unlike the law, ethical norms are not usually captured in black and white. There are different approaches to ethics (e.g. consequentialist versus deontological), and views about ethical behaviour with regard to armed conflict and international security may also differ from one country to another and may vary over time. However, some points of consensus have emerged as reflected by the 11 guiding principles that the Group of Governmental Experts on emerging technologies in the area of LAWS (GGE on LAWS) adopted in the framework of the CCW. One of the key principles is that humans should remain responsible for decisions on deployment of weapons and the use of force. As a result, technology should be designed in a way that (*a*) allows humans to exercise moral agency and oversight, and (*b*) prevents the creation of a gap in accountability by diffusing responsibility—that is, it should avoid causing a situation where it is difficult to assign responsibility when something goes wrong.[45]

Third, safety should be a central consideration to reduce the risk of misuse and unintended use of high-risk AI systems, in reference to the robustness component identified by the AI HLEG. High-risk systems include safety-critical (or life-critical) systems, such as cars, and systems that have a major impact on the functioning of societies and people's well-being such as power grids. This means ensuring that systems are robust in their design and do not have flaws that might, for example, make them more vulnerable to cyberattacks or adversarial attacks by malevolent actors.

Providing the knowledge needed

As outlined above, the three components (or pillars) provide a general framework for identifying the knowledge that AI actors might need to evaluate the outcome of their work in the context of international peace and security. The critical issue, however, is to ensure that AI actors obtain that knowledge. This could be done through various activities, including awareness raising, education and training.

Awareness raising can, for instance, take the form of initiatives, led by states or international organizations, that seek to inform the community of AI researchers and engineers. This can be achieved through publications targeted at entire sectors (e.g. following the model of the UN Guiding Principles on Business and Human Rights and the Organisation for Economic Co-operation and Development, OECD, Guidelines for Multinational Enterprises and Due Diligence Guidance for Responsible Business Conduct) or through education and training programmes targeted directly

[43] High-level Expert Group on Artificial Intelligence (AI HLEG), *Ethics Guidelines for Trustworthy AI* (European Commission: Brussels, Apr. 2019).

[44] For further discussion on the legal foundation of what constitutes responsible military use of AI see Boulanin, V., Goussac, N., Bruun, L. and Richards, L., *Responsible Military Use of Artificial Intelligence: Can the European Union Lead the Way in Developing Best Practice* (SIPRI: Stockholm, Nov. 2020); and Goussac, N., Bruun, L. and Boulanin, V., *International Humanitarian Law, Autonomous Weapons and Human Control,* (SIPRI: Stockholm, forthcoming 2021).

[45] Boulanin et al., *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control* (note 1); and Human Rights Watch, 'Losing humanity: The case against killer robots', Video, 19 Nov. 2012.

at companies, research institutions and universities.[46] National and regional export control systems are already engaged in activities with academia and the private sector that aim to raise awareness of and sensitivity to strategic implications, possible military end-uses and the threat landscape for international peace and security (see chapter 4).

One specific example of how education and training activities on RRI in AI could be conducted is a pilot initiative implemented by the UN Office for Disarmament Affairs (UNODA) in response to a call from the UN Secretary-General.[47] The initiative aims to sensitize academia and the private sector in the Asia–Pacific region to a 'holistic approach to responsible innovation in science and technology in the context of international peace and security'. What is remarkable about this initiative is its methodology. It focuses its efforts on raising awareness among university students in AI and related disciplines, using interactive, scenario-based exercises that encourage the students to reflect on the second and third order effects of the technology they are developing. Aware of the disincentives for private sector actors to engage directly in purely disarmament and non-proliferation activities, UNODA found that it is easier and more effective over the long term to educate engineers on international peace and security challenges related to the technology they are developing before they enter the market rather than after. The hope is that they will carry insights with them throughout their careers.

Awareness raising and education could also be implemented through the participation of international security, military or arms control and export control experts in existing forums where RRI in AI is discussed. The newly created European AI Alliance, which intends to organize regular assemblies with interested stakeholders, could be one example where such experts could directly engage with the wider AI community.[48] The participation of international security, military or arms control and export control experts in private sector-led RRI initiatives will probably be more difficult to achieve, but active outreach from states, international organizations like UNODA or civil society groups could help.

Awareness raising can also take place at the level of individual universities, research institutes and companies, through manuals, education, and training programmes targeted at students and researchers. These could be developed and provided by personnel working with professors from social science faculties, with ethics review boards or with internal compliance systems. Here again, interested states, international organizations like UNODA or civil society groups could help in the process by assisting universities, research institutes and companies to set up training programmes or provide access to experts or knowledge.

The means for implementing RRI

The implementation of RRI also requires that relevant stakeholders have relevant practical means to evaluate both outcomes and options in terms of international peace

[46] United Nations, Office of the High Commissioner for Human Rights, *Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework* (United Nations: Geneva, 2011); Organisation for Economic Co-operation and Development (OECD), 'OECD guidelines for multinational enterprises', 2011; and OECD, *OECD Due Diligence Guidance for Responsible Business Conduct* (OECD: Paris, 2018).

[47] The UN Secretary-General pledged to 'engage and work with scientists, engineers and industry to encourage responsible innovation of science and technology, to ensure its application for peaceful purposes, as well as the responsible dissemination of knowledge'. Under action item 28, the Secretary-General called on the UN Office for Disarmament Affairs to work with partners from civil society and academia to try to bring a holistic approach to responsible innovation in science and technology in the context of international peace and security. For further detail see UN Office for Disarmament Affairs (note 6).

[48] European Commission, 'The European AI Alliance', updated 10 Aug. 2020.

and security needs. Means refers both to the criteria that actors can use to guide their evaluation and to the mechanisms for carrying it out.

Various resources including handbooks on legal obligations, ethical and safety principles as endorsed by the university, company or the larger AI community (see discussion on existing guidelines in chapter 4) and risk assessment templates can all provide criteria. These documents may cover elements that could help researchers and engineers to evaluate legal, ethical, societal and safety aspects that specifically relate to international peace and security considerations. One concrete example is the self-assessment guide that was prepared by the European Defence Agency in 2019 to help applicants to the preparatory action on defence research to understand the international law and export control regulations with which they would need to comply.[49]

Methodologies for technology impact assessments and risk forecasting are other resources that could be relevant to the practice of RRI. There are already proven methods that the AI community could use to evaluate the potential impact of AI innovations on international peace and security: these include expert elicitation, scenario planning and formal modelling exercises.[50]

These documents and methodologies can be used in different settings. Indeed, RRI implementation structures and mechanisms come in different forms. These could be ad hoc or permanent multi-stakeholder structures that allow representatives from research, academia, the private sector and government with expertise in different disciplines to discuss or monitor desirable and undesirable outcomes of AI innovation on peace and security. Structures already exist where such deliberations could take place (see chapter 4). At the level of companies and universities, key implementation structures would include ethical review boards or internal compliance systems. These structures could, in fact, be linked. The internal compliance systems of companies and particularly the equivalent policies and procedures of universities and research institutes serve as systems that can integrate export control compliance, risk assessment and ethical and normative review mechanisms. Having integrated compliance and ethical review systems could help to link AI-related considerations in the ethical sphere with those that present legal obligations.[51]

Identifying possible outcomes

Ultimately, RRI should result in AI actors adopting functional requirements for or restrictions on the design, development and diffusion of, and trade in, new research, products and services that will limit the risks associated with their use. With regard to the risks posed by AI to international peace and security, RRI may lead to a variety of concrete measures.

In relation to the design and development of AI for instance, the adoption of high standards for reliability and predictability for a specific system or type of system (e.g. navigation systems in autonomous vehicles) could mitigate design-induced risks—such as system failure—that could cause harm. Researchers and engineers could determine these standards at the institutional level or at the sector level through industry standards.

In the case of the diffusion of AI, self-restriction in the dissemination of knowledge (e.g. the publication of algorithms or training data) or the adoption of technical measures—such as remote switches—that would allow the manufacturer to maintain

[49] European Defence Agency, 'Preparatory action on defence research (PADR) programme: Guidance on how to complete your self-assessment on "ethics, legal and societal aspects (ELSA)"', 19 Mar. 2019.

[50] Brundage (note 38).

[51] For more on this topic see chapter 4 (pp. 28–29) in this report.

some control could help to mitigate diffusion-induced risks, such as access to sensitive technology by irresponsible actors.

In terms of the use of AI, risk of misuse can also be limited through design decisions. One concrete example from the robotics industry was the decision of DJI, a Chinese producer of unmanned aerial vehicles (drones), to introduce by design a no-fly zone restriction in the control software of its products to limit the risk of this technology being used in conflict areas or near critical infrastructures.[52]

At a more general level, RRI processes could generate knowledge that might be useful for arms control discussions on the military use of AI. Principles, standards and technical measures that AI actors might adopt to promote responsible development, diffusion and use of AI could form a baseline for discussions between states on risk reduction measures with regard to the military use of AI.

[52] DJI, 'Fly safe: geo zone map', [n.d.].

4. Building on existing efforts to promote responsible research and innovation in AI

Responsible AI research and innovation is not a novel idea.[53] Therefore, the knowledge and means necessary to operationalize RRI in AI for international peace and security are already available—to some extent. This chapter maps out and explores the connections between the existing efforts focused on providing researchers, academia and industry with the knowledge and means to address risks associated with the development, diffusion and use of AI technology. The chapter aims to identify concrete elements that can be built on to make RRI in AI an effective upstream aspect of arms control. It starts with an overview of existing responsible AI initiatives that, for the most part, are intended to generate ethical and safety guidelines for the development and use of AI systems. It then discusses how RRI could complement export controls and internal compliance programmes (ICPs) and how these could help actors in research, academia and industry to act responsibly when engaging in the diffusion of AI technology.

I. Building on existing responsible AI initiatives

Responsible AI initiatives

In recent years a growing number of actors from academia, industry and government have voiced concerns about the need to proactively mitigate the negative impact that AI technology could have on society, the economy, democracy and human life. These actors have included scientists such as Stephen Hawking, entrepreneurs such as Elon Musk and Bill Gates, and politicians such as Canadian Prime Minister Justin Trudeau and French President Emmanuel Macron.[54] This groundswell of opinion led to multiple initiatives aimed at promoting RRI in AI. These initiatives have taken many forms and produced different outcomes (e.g. guidelines, codes of conduct, expert forums etc.). However, most often they have led to the introduction of ethical and safety principles and guidelines for responsible development, diffusion and use of AI technology (see box 4.1). According to Algorithm Watch—a non-governmental organization that maintains a database of existing initiatives—academia, industry and governments alike have published a total of at least 160 documents of this type.[55] These initiatives do not necessarily label themselves RRI explicitly, nor claim to have the ambition of implementing responsible innovation as defined in chapter 3. For the purpose of this report they will therefore be generally referred to as 'responsible AI initiatives'.

Challenges and opportunities

Existing responsible AI initiatives share many commonalities, notably in terms of area of focus (civilian use of AI), format (expert forums) and outcome (ethical and safety

[53] Dignum, V., *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (Springer International: Cham, 2019), pp. 47–69; and Brundage (note 38).

[54] Cellan-Jones, R., 'Stephen Hawking warns artificial intelligence could end mankind', BBC News, 2 Dec. 2014; Piper, K., 'Why Elon Musk fears artificial intelligence', Vox, 2 Nov. 2018; Piper, K., 'Bill Gates: AI is like "nuclear weapons and nuclear energy" in danger and promise', Vox, 20 Mar. 2019; and Knight, W., 'Canada and France plan an international panel to assess AI's dangers', *MIT Technology Review*, 7 Dec. 2018.

[55] This was correct as of Apr. 2020, but the actual number may be greater given that the database is dependent upon contributors and Algorithm Watch. For details see Haas, L. and Gießler, S., 'In the realm of paper tigers: Exploring the failings of AI ethics guidelines', Algorithm Watch, 28 Apr. 2020. For a meta-analysis of 84 different ethical principles or guidelines, along with a detailed overview and analysis of their content, see Jobin, A., Ienca, M. and Vayena, E., 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, vol. 1 (Sep. 2019), pp. 389–99.

Box 4.1. Notable responsible AI initiatives

Notable responsible artificial intelligence (AI) initiatives include the following:

- The Institute of Electrical and Electronics Engineers (IEEE) global initiative on ethics for autonomous and intelligent systems—an initiative by the AI engineering community.[a]
- The Partnership on AI—a joint initiative between for-profit technology companies, representatives of civil society and academic and research institutions, including leading companies such as Google and Facebook.[b]
- The High-level Expert Group on AI (AI HLEG)—an initiative established by the European Commission.[c]
- The Global Partnership on AI—an initiative proposed by the French and Canadian governments that has 15 founding members, with the Organisation for Economic Co-operation and Development (OECD) hosting its secretariat in Paris. The OECD itself has developed its own AI principles, which were originally adopted by 42 different countries.[d]
- The United Nations Secretary-General's Roadmap on Digital Cooperation—an initiative that has raised some issues related to responsible research and innovation (RRI) in AI, as documented in Recommendation 3C.[e]

[a] Institute of Electrical and Electronics Engineers (IEEE), *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* (IEEE: 2019).
[b] See the Partnership on AI website.
[c] European Commission, 'High-Level expert group on artificial intelligence', 9 July 2020.
[d] The founding members are Australia, Canada, the European Union, France, Germany, India, Italy, Japan, Korea (Republic of), Mexico, New Zealand, Singapore, Slovenia, the United Kingdom and the United States. For more information see OECD, 'OECD to host secretariat of new Global Partnership on Artificial Intelligence', 15 June 2020; and OECD, 'Forty-two countries adopt new OECD principles on artificial intelligence', 22 May 2019.
[e] United Nations, 'Secretary-General's High-level Panel on Digital Cooperation', [n.d.].

principles or guidelines). These commonalities create challenges but they also provide opportunities for supporting the use of RRI in AI to achieve arms control objectives.

Challenges

A first notable challenge is that existing initiatives focus almost exclusively on risks in the civilian realm and have produced little knowledge about the risks posed by AI to international peace and security. They typically aim to prevent and mitigate the negative impact of the use of AI in healthcare, transportation, industry, public government or judiciary systems. The use of AI for military purposes is usually not addressed. The few initiatives that do address the topic of military use of AI were drafted by military actors. These include (*a*) the ethical principles for AI developed by the US Department of Defense (DOD); (*b*) the TrUE AI approach of the French arms company Thales, which aims to demonstrate the company's commitment to producing transparent, understandable and explainable AI systems; and (*c*) the ethical review committee set up by Airbus to inform the development of its Future Combat Air System (FCAS).[56] There are perhaps three key reasons for the strong civilian focus and lack of consideration for military issues in existing responsible AI initiatives.[57] First, the majority of initiatives have been established by actors that research and develop AI for civilian end-uses. It is therefore not surprising that they prioritize challenges in the civilian sector, especially given that AI already raises many important issues in this area, such as bias and discrimination, privacy issues, and potential issues related to the denial of individual rights.

[56] US Department of Defense, Defense Innovation Board, 'AI principles: Recommendations on the ethical use of artificial intelligence by the Department of Defense Defense Innovation Board', 2019, p. 4; Lopez, C. T., 'DOD adopts 5 principles of artificial intelligence ethics', US Department of Defense, 25 Feb. 2020; Thales Group, 'The Thales true AI approach to be unveiled at Paris air show 2019', 6 July 2019; and Airbus, 'Future Air Combat Systems, Airbus and Fraunhofer FKIE create expert panel on the responsible use of new technologies', 14 May 2020.
[57] These points are drawn from the authors' interviews with experts and industry representatives conducted under the Chatham House Rule.

Second, these very same actors might lack knowledge about possible military end-uses of civilian AI innovation and about the strategic and humanitarian consequences of adoption and use of AI by the military. Third, private companies that are behind a number of major initiatives may have concerns that expressing views on military use of AI could create bad publicity and negatively impact sales or funding. In some countries, notably in the USA, the defence ministry is also a key source of funding for universities and companies, which might discourage researchers and engineers from taking a stance in public debates on the use of AI for military purposes.

Another notable challenge is that the majority of the existing initiatives focus almost exclusively on the responsible development of AI. Risks related to AI diffusion are not commonly considered, although there have been some efforts to move into this area over the past two years. Notably, in 2019 the research company OpenAI initiated a discussion on a 'responsible publication norm' following its decision to limit information around the release of GPT-2, a natural language processing model that can generate or summarize coherent text, provide machine translation and answer questions.[58] OpenAI justified its decision not to release the key elements of the data set and AI training code because of concerns that the program could be used by malevolent actors to generate fake news, spam content or impersonate people online.[59] OpenAI's decision led to various discussions about the tension between open research and precautionary concerns. The Partnership on AI organized a public discussion that led to specific recommendations on the need not only to conduct standard risk assessment processes, but also to exercise precaution during research design and scoping.[60]

In terms of the format and outcome of existing initiatives, there are also two notable challenges. First, these initiatives have so far been limited to the introduction of ethical and safety principles and guidelines. They commonly lack implementation mechanisms that would provide practical means for AI actors to engage in RRI throughout a technology's life cycle. This can be partly explained by the fact that these initiatives are often very new—therefore, it is hard to know if, how and to what extent they have been implemented so far. Nevertheless, some of these efforts have already had a noticeable impact. For example, in 2020, two years after the publication of its AI principles, Google implemented a number of measures for responsible AI development. These measures included (*a*) making technical ethics training available to all its employees; (*b*) issuing publications that address technical approaches to fairness, safety, privacy and accountability; (*c*) enlisting the non-profit organization Business for Social Responsibility to conduct a formal human rights assessment based on the UN's Guiding Principles on Business and Human Rights to feed into the design of a product; and (*d*) conducting internal and external engagement at various levels.[61]

The AI HLEG is another example of an initiative that has made notable progress in terms of implementation. For example, it has published principles for ethical and robust design of AI systems and was instrumental in the creation of the European AI Alliance.[62] In addition, it has developed an AI risk assessment list for researchers to use to evaluate the various kinds of risk that may be posed by the systems they

[58] With the release of GPT-3, OpenAI took a different approach for commercial reasons but also stated that, because 'it is hard to predict the downstream use cases of [its] models, it feels inherently safer to release them via an API [application programming interface] and broaden access over time, rather than release an open source model where access cannot be adjusted if it turns out to have harmful applications'. For further detail see OpenAI, 'OpenAI API', 11 June 2020.

[59] Crootof, R., 'Artificial intelligence research needs responsible publication norms', Lawfare Blog, 24 Oct. 2019.

[60] Leibowicz, C., Adler, S. and Eckersley, P., 'When is it appropriate to publish high-stakes AI research?', Partnership on AI, 2 Apr. 2019.

[61] Walker, K. and Dean, J., 'An update on our work on AI and responsible innovation', Google Blog, 9 July 2020. See also United Nations, Office of the High Commissioner for Human Rights (note 46).

[62] European Commission (note 48). On the principles for ethical and robust design of AI systems see High-level Expert Group on Artificial Intelligence (note 43).

Figure 4.1. Frequently cited principles for responsible AI

Source: Jobin, A., Ienca, M. and Vayena, E., 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, vol. 1 (Sep. 2019), pp. 389–99.

design.[63] Some other organizations, such as AIGlobal, also aim to provide AI actors with resources and concrete tools to engage in responsible AI development.[64]

A second and perhaps bigger challenge is the fact that these initiatives are often developed and adopted within specific silos. As mentioned previously, at least 160 responsible AI guidelines have been produced. These initiatives typically tend to be adopted by particular clusters of actors in a region, country, industry, research community or company. While these documents often recommend more or less the same set of high-level principles for responsible AI—including transparency, justice and fairness, responsibility and privacy—there is little interactivity between them (see figure 4.1). There are perhaps both positives and negatives to this fragmented approach. On the one hand, it could allow not only for RRI practices that are tailor-made to a specific context, but also for a truly bottom-up emergence of widely recognized norms for responsible AI in the long run. On the other hand, it could be argued that such fragmentation limits the effectiveness, and hence the value of, RRI as a self-governance mechanism. There is notably a risk that actors involved in these various initiatives focus only on issues and adopt self-restraint measures that align with their specific interests. The fact that most existing responsible AI initiatives are

[63] High-level Expert Group on Artificial Intelligence (note 43).
[64] AI Global, 'About', [n.d.].

driven by civilian actors and do not include the participation of military arms control stakeholders is problematic from that standpoint. Indeed, it would be valuable if existing responsible AI initiatives were to take into consideration the potential risks associated with the use of AI in the military sector in a more explicit manner. While this would help AI actors to understand and limit the possible second and third order effects of their work on international peace and security, it would be difficult to achieve without greater interactivity with experts from government, industry and civil society that have knowledge about military use of AI and the potential risks generated by such use. It is unlikely that civilian actors will realize and act on the near- or longer-term implications of their research and innovation for peace and security on their own—for both conceptual and commercial reasons.

At the conceptual level, it might be relatively easy for civilian actors to realize that their work on, for instance, computer vision systems (such as facial recognition) could end up helping with military ISR systems or AWS and affect the way militaries select targets. However, it might be more difficult for civilian actors to foresee that design choices related to, for instance, the navigation control of autonomous underwater systems could have implications for conflict escalation and nuclear risk.[65] In this case, it might be hard for them to recognize ownership of the problem when the logical chain is so long and hypothetical.

At the commercial level, there are limits to what can be expected from civilian actors in terms of voluntary efforts and self-restraint, given that for some of them—especially civilian companies—engaging with military-related issues could potentially lead to bad publicity and hence a commercial risk. Companies in particular might need to be motivated to engage on these issues through economic incentives. One idea that has been suggested would be to create a civil liability mechanism that would engage the responsibility of companies in tech-related harm.[66] This could push civilian companies to discuss, and take proactive measures to mitigate, the risks related to the potential military end-use of their technology.

Opportunities

Despite the above-mentioned challenges, there are nonetheless elements of commonalities between the various responsible AI initiatives that provide some opportunities for using RRI in AI as a way to achieve arms control objectives. This section explores two key opportunities.

First, some of the existing initiatives have, through the production of ethical and safety guidelines, set standards for the responsible development of AI that are relevant for both the civilian and the military sectors. As previously mentioned, existing initiatives have resulted in documents that in essence focus on the same set of high-level principles. In this regard, it is also worth noting that some military-led initiatives have proposed principles that are very similar to those in civilian responsible AI initiatives. The US DOD, for example, has stated that AI should be in line with the following principles:

1. *Governable*. AI should be designed in a way that would allow users to maintain agency and oversight, and thereby detect and avoid unintended consequences, as demanded by legal and ethical frameworks. This should include the possibility to override or deactivate the system.

[65] On this scenario see Boulanin et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (note 1).

[66] Crootof, R., 'War torts: Accountability for autonomous weapons', *University of Pennsylvania Law Review*, vol. 164, no. 6 (May 2016).

2. *Equitable*. AI should be designed in a way that minimizes the risk of unintended bias.

3. *Explainable*. AI should be designed in a way that allows relevant developers and users to adequately understand how the technology works (transparency) and be able to trace back sources of problems when something goes wrong (traceability). This can be achieved through the design of transparent and auditable methodologies, data sources, design procedures and documents.

4. *Safe*. AI should be reliable and designed in a way that mitigates the risk of unintended consequences when confronted with situations that were not foreseen at the programming stage or when challenged by adversarial or malicious actors. Both the requirement of governability and explainability derive from this technical foundation.[67]

These principles are all very relevant for addressing the humanitarian and strategic challenges that might result from the way AI technology is designed and used. Notably, they can help not only to assuage concerns around possible accountability gaps in the case of IHL violations or accidents related to the use of military AI, but also to reduce the risk of incidents that could lead to accidental or inadvertent escalation in conflict.[68]

Second, the processes that existing responsible AI initiatives intend to develop—albeit with a focus on civilian end-uses—could directly or indirectly contribute to responsible development and use of AI in the military sphere. The risk assessment processes or testing and evaluation standards and procedures adopted by the civilian sector to address AI challenges will set benchmarks. Stakeholders in the military sector—researchers, defence industry and military institutions alike—will have to take these benchmarks into consideration and may need to build on them, for two key reasons. First, the critical challenges associated with the design of AI systems, such as transparency, understandability, explainability and reliability, are common to civilian and military AI innovations—albeit they generally follow different pathways.[69] Second, the civilian sector, as the driving force of innovation in AI, already influences how AI military technology is developed and used. Civilian applications (within a given category of product) of AI are generally easier and cheaper to design and deploy than military applications because they do not need to follow as strict a standard for safety, security and reliability as their military equivalents.[70] One consequence is the emergence (or reinforcement, depending on the case) of a pattern in the military sector whereby defence engineers innovate through adaptation: rather than developing military solutions from scratch, engineers in the defence sector sometimes attempt to modify a civilian innovation for military purposes. In other words, civilian technology can end up forming the basis for military applications.

From that standpoint, the standards and measures that the responsible AI initiatives recommend for the development of civilian applications could have indirect yet positive outcomes on how military AI is developed. They could help to reduce some of the development-induced risks that are associated with the military use of AI. The norms and processes that the responsible AI initiatives might recommend to limit the misuse of AI research and innovation while maintaining openness of AI developments—like

[67] Lopez (note 56).

[68] Crootof (note 66).

[69] In the early stage of R&D, civilian and military innovations may share fairly similar trajectories because at the fundamental level they use the same development methods and principles. In the later phase of development, military applications usually need to follow a stricter standard for safety, security and reliability than their civilian equivalents. Boulanin and Verbruggen (note 8).

[70] Boulanin and Verbruggen (note 8).

OpenAI did in the case of GPT-2 and more recently GPT-3—could also help to prevent the diffusion of risks associated with dual-use AI innovations.[71]

II. Building on export controls and compliance systems

Export control regulations and internal compliance programmes in academia, research institutes and the private sector

Export controls are the policies states put in place to govern the movement of military equipment and dual-use goods and technology. States adopt export controls to implement a variety of norms and obligations, often found in arms control treaties with a non-proliferation focus, such as the 1968 Treaty on the Non-Proliferation of Nuclear Weapons (Non-Proliferation Treaty) and the 2013 Arms Trade Treaty. They also implement some of the norms and provisions in IHL, human rights law, UN Security Council resolutions and the multilateral export control regimes.[72] As states' international commitments and control standards have expanded in these areas, export controls have widened the scope of transactions to which they apply—for example, to cover brokering and academic publishing—and have put more responsibility on companies, research institutes and universities to exercise due diligence.[73] In recent years export controls have also moved more firmly into the preventive area as coverage of emerging technologies has grown and risk assessment obligations have increasingly become part of the responsibilities of exporters. In this way, RRI in AI can build on export controls and the ICPs that companies, research institutes and universities put in place, particularly with regard to knowledge on applicable legal restrictions and potential diffusion risks.

Discussions are intensifying among the export control community about dual-use applications of AI technology that could be integrated in conventional weapons and delivery systems—and potentially nuclear weapon-related systems—as well as in military logistics, infrastructure and decision-making systems. Export controls already apply to AI-enabled military items and—directly and indirectly—a range of dual-use AI-enabling or -related hardware, software and technology.[74] For example, the Wassenaar Arrangement (WA), which is one of the multilateral export control regimes, already controls some technologies covered by the AI category, including neural network technologies.[75] Controls on software and technology are more indirect and cover any software and technology 'designed or modified' for the 'development, production and maintenance' of equipment, software and materials covered by the WA's Munitions List.[76] Multilateral (within the WA), regional (within the EU) and national discussions on expanding the scope of export controls to more explicitly

[71] Bostrom, N., 'Strategic implications of openness in AI development', *Global Policy* (2017); and Crootof (note 59).

[72] For an overview and discussion of the export control regimes see Joyner, D. H. (ed.), *Non-proliferation Export Controls: Origins, Challenges, and Proposals for Strengthening* (Ashgate: Aldershot, 2006).

[73] Bauer et al., *Challenges and Good Practices in the Implementation of the EU's Arms and Dual-use Export Controls: A Cross-sector Analysis* (SIPRI: Stockholm, July 2017), pp. 1–2.

[74] Viski, A. et al., *Artificial Intelligence and Strategic Trade Controls*, Technical Report (Strategic Trade Research Institute and Center for International and Security Studies at Maryland: Washington, DC, and College Park, MD, June 2020), pp. 44–45; Rasser, M. et al., *The American AI Century: A Blueprint for Action* (Center for a New American Security: Washington, DC, Dec. 2019); International Panel on the Regulation of Autonomous Weapons (iPRAW), 'LAWS and export control regimes: Fit for purpose?', iPRAW Working Paper, Apr. 2020; Flynn, C., 'Recommendations on export controls for artificial intelligence', Center for Security and Emerging Technology (CSET), *CSET Issue Brief*, Feb. 2020; and Stanley-Lockman, Z., 'Why the sky is not falling: The diffusion of artificial intelligence', *Eurasia Review*, 26 June 2019.

[75] Thomsen, II, R. C., 'Artificial intelligence and export controls: Conceivable, but counterproductive?', *Journal of Internet Law*, vol. 22, no. 5 (Nov. 2018), pp. 1, 15–24.

[76] Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-use Goods and Technologies (Wassenaar Arrangement), 'List of dual-use goods and technologies and munitions list', vol. II, 5 Dec. 2019.

cover specific contemporary AI technologies are currently being considered by some states—driven in large part by US export control reform efforts.[77]

The ICPs already established by actors affected by export controls can often be important tools to govern transfers of AI technology and might also contribute to RRI in AI. An ICP is a set of means and procedures that an entity seeking to comply with export controls—and in many cases other legal obligations and internal policies such as codes of conduct—puts in place to ensure that 'it is completing legal transactions, obeying the regulations enacted by the government, and fulfilling company export policies'.[78] An ICP thus establishes key internal oversight functions and allocates personnel with legal expertise to work with employees, including researchers and developers, to ensure that their actions and the actions of the company, research institute or university do not violate any applicable laws and internal policies. ICPs set up systems that raise awareness and provide project leaders and researchers with tools to identify sensitive research, conduct risk assessments and ensure that they apply for any required licences when transferring, sharing, making available, or publishing their research, technology and know-how. ICPs routinely include regular training, internal guidance documents, red flag indicators and information technology (IT) systems, such as digital access management systems, in order to prevent inadvertent violations of export controls and to ensure proper record keeping.[79]

Beyond compliance with export control regulations, risk assessments often also examine transactions for reputational risks. Such risks generally do not result from illegal activities; in most cases they are rooted in ethical or normative concerns that may result in negative perceptions and a backlash. As such, these risk assessments can be informed by RRI processes and implement guidance on responsible practices. In terms of AI technology applications, facial recognition technology is an example of an area where companies—in the absence of strong legal restrictions—have been self-restricting based on reputational and normative considerations and actually demanding additional regulation.[80]

Challenges and opportunities

This section aims to show that while there are challenges associated with the effectiveness of export controls and compliance systems, there are also opportunities for strengthening export controls and leveraging ICPs to implement RRI in AI in the pursuit of arms control objectives.

Challenges

The compliance obligations that export controls place on research institutions, universities and companies already require them to have mechanisms and processes in place to behave responsibly in the diffusion of AI technology. In practice, however, export controls and internal compliance systems face many challenges, potentially rendering them less effective, including with regard to enabling RRI. There are at least three key challenges related to the regulatory system of export controls and at least five related to the use of ICPs to diffuse AI technology responsibly.

[77] Barkin, N., 'Export controls and the US–China tech war: Policy challenges for Europe', MERICS China Monitor, Perspectives, 18 Mar. 2020, p. 7; and Griffiths, P., Head of Secretariat, Wassenaar Arrangement, 'The proliferation threat landscape in 2017: Mounting dangers? WMD/military proliferation trends and emerging technologies of concern', Remarks at the 2017 Export Control Forum, Brussels, 19 Dec. 2017.

[78] Institute for Science and International Security, 'Key elements of an effective export control system', 2003.

[79] Bauer (note 73), pp. 41–42.

[80] Shepardson, D., 'IBM says US should adopt new export controls on facial recognition systems', Reuters, 11 Sep. 2020; and IBM, 'A precision regulation approach to controlling facial recognition technology exports', IBM THINKPolicy Blog, 11 Sep. 2020.

One important regulatory challenge is the lack of clarity on the coverage and application of export controls to AI. This makes consideration of, and compliance with, the legal provisions as part of RRI difficult. Moreover, agreeing on changes to the control lists in the multilateral export control regimes requires consensus among the participating states and is particularly burdensome in the case of emerging technologies, which develop and change key parameters quickly.[81] This process can be even more problematic for enabling technologies such as AI that cut across the traditional divisions of the multilateral export control regimes.[82]

A second regulatory challenge relates to the lack of harmonization of the interpretation of the criteria used in licensing decisions.[83] National governments use criteria (i.e. sets of principles or considerations) for determining whether to grant or deny an export licence.[84] If and how these criteria apply to the potential military applications of AI remains ambiguous, and none of the accompanying national or multilateral guidance material includes any specific considerations on the destabilization potential of AI-enabled military systems. In addition, it is unclear whether there are any established good practices by states for applying export controls to AI. Some decisions that academia, research institutes and companies take will have to be informed by other considerations and potentially be self-restrictive.

The third key regulatory challenge is that most transfers of AI technology are digital transfers or are transfers of knowledge and technology in an intangible form or by intangible means.[85] Such intangible transfers of technology (ITT) are difficult to detect, investigate and prosecute. Adherence to the controls on ITT and the maintenance of adequate record keeping of intangible transfers are also challenging from a compliance perspective.[86] Only a few countries have so far adopted special audit procedures for ITT that use digital forensics techniques and check the records of all digital transfers, or have exchanged good practices among specialized prosecutors.[87]

ICPs that universities, research institutes and companies have established to comply with export controls already provide a concrete framework through which AI innovators can ensure that they act responsibly when engaging in the diffusion of AI technology. However, establishing and maintaining such a system comes with many challenges. Five are explored here.

First, the burden and the resources available to shoulder compliance obligations vary significantly, particularly between different sectors of industry and actors in research and academia, and according to the size of the entity. Most larger companies and research institutes, as well as a growing number of universities, have ICPs in place. However, many smaller companies and institutes engaged in R&D, especially from sectors that are commonly little affected by export controls—such as the AI sector—may not have ICPs and may not allocate specific personnel to compliance functions.

[81] Brockmann, K., 'Drafting, implementing, and complying with export controls: The challenge presented by emerging technologies', *Strategic Trade Review*, vol. 4, no. 6 (spring/summer 2018).

[82] Brockmann, K., *Challenges to Multilateral Export Controls: The Case for Inter-Regime Dialogue and Coordination* (SIPRI: Stockholm, Dec. 2019).

[83] Leung, J., Fischer, S. and Dafoe, A., 'Export controls in the age of AI', War on the Rocks, 28 Aug. 2019.

[84] Council of the European Union, Council Common Position 2008/944/CFSP of 8 Dec. 2008 defining common rules governing control of exports of military technology and equipment, *Official Journal of the European Union*, L335, 8 Dec. 2008. Amended by Council Decision (CFSP) 2019/1560 of 16 Sep. 2019, *Official Journal of the European Union*, L239, 17 Sep. 2019; and Wassenaar Arrangement, 'Elements for objective analysis and advice concerning potentially destabilising accumulations of conventional weapons, as adopted in 1998 and amended by the plenary in 2004 and 2011', 2011.

[85] Leung, Fischer and Dafoe (note 83).

[86] Bromley, M. and Maletta, G., *The Challenge of Software and Technology Transfers to Non-proliferation Efforts: Implementing and Complying with Export Controls* (SIPRI: Stockholm, Apr. 2018); and Bauer, S. and Bromley, M., *Detecting, Investigating and Prosecuting Export Control Violations: European Perspectives on Key Challenges and Good Practices* (SIPRI: Stockholm, Dec. 2019).

[87] Bauer and Bromley (note 86).

Setting up and managing an ICP can incur significant costs, including for employing staff to run the ICP, training scientists and research staff, and acquiring screening software and other IT tools.[88] ICPs therefore need to be tailored and adapted to the structure, size and sector of an entity and 'integrated into standard procedures and business practices'.[89]

Second, ICPs require in-reach and awareness raising, especially among scientists and research staff, who in many cases lack awareness of their responsibility to assess the potential risks associated with their AI research and its diffusion. They also often lack the knowledge about which tools they can use to facilitate conducting such risk assessments.[90] This can become an issue of particular importance with regard to other actors in the supply chain and applied research collaboration where the partners are largely only engaged in basic scientific research, but are required to comply with regulations and conduct risk assessments because they contribute to applied research.

Third, to date, no sector-specific guidance materials on compliance in the AI sector exist and thus there is a lack of common standards for compliance functions in this field. While the availability of guidance on export control compliance for research and academia has improved (particularly in the context of the EU) over the past few years, this guidance generally does not include examples from the AI sector or specific recommendations for AI-related risk assessments.[91] The classification of new products, as well as the technology developed and the know-how created during research and innovation, can often be difficult if it is unclear to what extent the product, technology or know-how may already be covered by dual-use or arms export controls, and if other regulations require specific actions or precautions.[92]

Fourth, obtaining a complete picture of the risks posed by transfers of AI technology and the potential misuse of this technology by other actors can be challenging because of the limited information that compliance officers in companies, research institutes and universities have available to them. While 'red flags', suspicious party lists and knowledge about illicit procurement tactics are more established for non-proliferation of conventional and chemical, biological and nuclear weapons, there is far less information available concerning AI.

Finally, the field of AI research has a very distinct open source and data sharing culture that translates into a lack of willingness to engage with restrictions that are often perceived as contravening academic freedoms. This means that there is a need to raise awareness in the wider AI sector and to create forums where compliance officers from AI companies and those involved in R&D can meet to discuss shared challenges and good practices.

Opportunities

Export control systems and related ICPs provide not only a means to govern the diffusion of AI technology, but also a way to help stakeholders to engage in RRI in AI. This section identifies nine opportunities to strengthen export controls and build on ICPs in implementing RRI in AI to achieve arms control objectives.

[88] Bauer (note 73), p. 2.

[89] Bauer (note 73), p. 41.

[90] Bauer (note 73), p. 28.

[91] See e.g. the ongoing public consultation on EU compliance guidance for research involving dual-use items and the German guidance documents for research and academia. European Commission, 'Targeted consultation on draft EU compliance guidance for research involving dual-use items', updated 2 Nov. 2020; German Federal Office for Economic Affairs and Export Control (BAFA), *Export Control in Science and Research* (BAFA: Eschborn, Feb. 2019); and BAFA, *Export Control and Academia: Manual* (BAFA: Eschborn, Feb. 2019).

[92] Head of export control compliance of a European research institute, Interview with the authors, 3 Sep. 2020; and Bauer (note 73), p. 33.

First, the use of multilateral export controls on AI—rather than unilateral controls—and the harmonization and clarification of the application of the relevant criteria could strengthen the regulations without creating significant adverse effects and competitive disadvantages. The process of updating export control lists and guidance increasingly involves public consultations and greater input from actors in industry, research and academia. This provides an opportunity for reflections on RRI in AI to be taken into consideration during the international discussions on changes to export controls. The reflections made in an RRI process informed by considerations of export controls and other regulations, as well as considerations of ethics and robustness aspects, could establish an important feedback loop.

Second, the 'catch-all' mechanism foreseen in most export control systems could also present an opportunity and be a useful tool to help to control emerging military applications of AI. This mechanism foregoes unnecessarily disruptive control on multipurpose technologies by 'creating a legal mechanism to prohibit exports in case of suspicion or knowledge of an undesirable end-use'. This means that it does not place a prohibition on a specific technology itself and thus should not disrupt innovation.[93]

Third, strengthening national capabilities to enforce export controls—particularly concerning ITT, which is key in the context of AI—could help to disincentivize non-compliance and would mean that actors that irresponsibly diffuse military AI technology are properly penalized. Strengthening the capabilities of national authorities to assist in the classification of goods and technologies or to provide guidance on this could help with compliance procedures and necessary risk assessments on AI technology throughout the innovation cycle of a technology.

A fourth opportunity is presented by the growing volume of guidance for and outreach to research and academia provided by export licensing authorities, which could be the basis for sector-specific guidance and outreach efforts to the AI sector. Whether through good practices or codified in a particular provision in guidance documents, explicit inclusion of considerations related to the destabilizing potential of AI-enabled military systems could then be included in RRI and cause stakeholders to undertake more responsible actions to avert this risk.

Fifth, ICPs provide key functions that help companies, research institutes and universities to obtain knowledge about legal provisions and procedures as well as systems to comply with them. Thus, they offer important opportunities for RRI to build on and use the functions and procedures that already exist. ICPs enable RRI processes to consider and analyse in more depth non-proliferation and arms control objectives, and international legal norms in risk assessments.

A sixth opportunity relates to the incentivization of the wider adoption of ICPs. In some states there are formal requirements for exporters to have an ICP with specific characteristics in place to apply for certain export licences for multiple transfers or trade facilitation mechanisms.[94] Making a certified ICP a requirement for access to such licences and mechanisms could help to incentivize the adoption of ICPs.

Seventh, despite the costs they incur, ICPs should be seen as an asset for companies, research institutes and universities as they allow them to derive benefits from access to simplified export procedures, and can facilitate the risk assessments required when taking responsible decisions in the development of AI technology, making funding applications or seeking export licence approval.[95] In addition, they help to reduce the

[93] Bauer, S., 'New technologies and armament: Rethinking arms control', *Clingendael Spectator*, 29 July 2020.
[94] Bauer (note 73), p. 3.
[95] Bauer (note 73), p. 2.

risk of inadvertent violations of regulations and of reputational damage. They are also an increasingly significant factor in attracting customers and investors.[96]

An eighth opportunity is provided by the complementarity of RRI with export control compliance. The lack of clarity and differences in national implementation with regard to the exemptions from export controls for basic scientific research and the freedom of research more generally continue to create challenges for compliance departments.[97] RRI could provide a strong complementary measure that helps stakeholders to assess such issues and potentially self-restrict if there are concerns beyond existing legal frameworks that demand this.

Finally, the processes and training programmes already in place in many compliance departments could potentially be expanded to incorporate or connect to ethical and robustness reviews, so that all three pillars of responsible AI are addressed. Formal and procedural connections between compliance departments and ethics boards—even if just to explain their approaches to questions and processes—could be a helpful step to link functions.[98] According to one compliance professional, this could be particularly helpful in applying additional scrutiny to and deciding on cases of transfers involving AI-enabled technologies, such as facial recognition technologies, that can facilitate state surveillance. While such a transfer might in all other ways be legal, consideration should be given, for example, to ethical aspects and whether it could potentially lead to human rights violations.[99]

III. Conclusions on synergies between responsible AI initiatives and export control compliance

In summary, this chapter explains that exploring synergies between different elements of responsible AI initiatives and export control compliance is important to make RRI an effective approach to improving the governance of risks posed by AI technologies for international peace and security. While consideration for legal compliance, ethics and robustness in the development process of AI technologies can reduce some risks, other aspects require responsible practices in the realm of export control to prevent possible risks associated with diffusion of dual-use AI technology. These two areas of intervention could be more closely connected, particularly in relation to the resources and means that they are, or could be, deploying to encourage AI actors in research, academia and the private sector to integrate peace and security considerations in their ongoing work and reflections on responsible development of AI technology. Guidelines, risk assessment methodologies and processes, awareness raising, education and training activities, ethical review boards, and internal compliance mechanisms could all be more closely linked. This could help AI actors to consider and implement in an integrated and holistic way their legal compliance obligations as well as the ethical and safety standards to which they are subscribed.

[96] Compliance professional in a major international company developing AI, Interview with the authors, 27 Aug. 2020.

[97] Head of export control compliance of a European research institute (note 92).

[98] Compliance professional in a major international company developing AI (note 96).

[99] Compliance professional in a major international company developing AI (note 96).

5. Key findings and recommendations

This report has explored how the risks posed by the development, diffusion and military use of AI could be mitigated through the adoption and promotion of RRI as an upstream approach to arms control. This chapter summarizes the key findings and provides some recommendations.

I. Key findings

The development, diffusion and adoption of military and dual-use applications of AI is not inevitable; rather it is a choice, one that must be made with due mitigation of risks.

The arms control community is currently considering the role it can play in ensuring that the risks posed by AI technologies are addressed. It is still debating to what extent the standard tools of arms control can mitigate the humanitarian and strategic risks posed by the military use of AI. The fact that such use hides a complex technological reality makes the discussion on the topic challenging. AI is an enabling technology that transcends the technology-centric silos in which arms control processes usually operate. It also requires a level of technical expertise that states—as the central actors in arms control processes—might not be able to mobilize sufficiently and quickly enough to understand and react to rapid developments in this area. In addition, AI has become the object of great power competition, which adds geopolitical challenges to the pursuit of an arms control response to the risks related to military use of AI.

In this context, the report found that RRI as an approach to technology governance could be useful for several reasons. First, it aims to involve all relevant stakeholders, particularly academia and industry, which have the technical understanding of the risks that may result from the development, diffusion and military use of AI technology. Second, it provides a governance framework for the early phase of R&D that arms control may not easily capture. Third, RRI is preventive and, by nature, iterative. It aims to identify risks and act upon them before they materialize. Moreover, it seeks to do so not just once but throughout the life cycle of technologies. Finally, because it does not necessarily aim to impose hard regulations, RRI is potentially a less politicized process than formalized arms control discussions. Like arms control, however, RRI also has its limitations. It is only one approach among others and lacks harmonized implementation and enforcement mechanisms.

At the same time, the principles and self-governance instruments that RRI creates could help the arms control community to make advances in its deliberations on the governance of the risks posed by AI. Notably, RRI processes could build on existing responsible AI initiatives, export controls and internal compliance systems.

Many of the initiatives launched in recent years have targeted the development of principles and mechanisms for RRI in AI. They typically do not address risks related to the military use of AI—although they clearly should, given the predominantly dual-use nature of AI innovation. Against this backdrop, the report explored ways through which existing RRI efforts on AI could mainstream international peace and security considerations. It found that there is a need to increase awareness about the second and third order effects of AI research and innovation, from both a humanitarian and a strategic standpoint. The report discussed how AI researchers and engineers could evaluate and limit the consequences of their work through a number of means. These could include (*a*) the implementation of very high ethical and safety standards; (*b*) the development of mechanisms and methodologies for technology impact assessments and foresight; (*c*) the design of fail-safe mechanisms; and (*d*) the application of precautionary measures in the publication of research findings. Universities, research

institutes and companies already diffuse AI technology in a responsible way by complying with obligations derived from export control regulations and conducting risk assessments required by funding organizations. ICPs also provide procedures, training and systems that help researchers and developers to comply with legal provisions. In the case of AI technology, the report found that it would be a good practice to connect such compliance systems with ethical review mechanisms and robustness checks to enable a comprehensive reflection on these aspects. Ultimately, RRI should lead to decisions in the innovation and commercialization processes that can help to prevent, or pre-emptively mitigate, risks associated with the development, diffusion and military use of AI.

II. Recommendations

In the light of these findings, this report makes key recommendations targeted at companies, research institutes and universities—as well as states and regional organizations—that already promote or could promote RRI as a valuable approach to govern the risks posed by the military use of AI.

Companies, research institutes and universities

Mainstream peace and security considerations into existing initiatives on responsible AI

Existing responsible AI initiatives should give greater consideration to the risks posed by the military use of AI. They should make the risk associated with military end-use of AI a theme in (*a*) existing discussion forums, (*b*) education and training activities, and (*c*) ethical reviews and risk assessment processes.

Connect responsible innovation mechanisms and internal compliance programmes

Compliance officers in companies, research institutes and universities should work more closely with ethics boards and similar internal oversight bodies. This could allow for a closer link between the risk assessments from their respective standpoints. Strengthening the connections between compliance programmes and ethics review mechanisms could also allow companies, research institutes and universities to take advantage of existing training and in-reach processes and improve the understanding of researchers and engineers of the peace and security risks they need to consider, particularly in the wider AI field.

States and regional organizations

Consider ways to consult with the AI sector in arms control discussions on AI

States and regional organizations, such as the EU, should find ways to facilitate greater engagement of the AI research community and industry in arms control discussions. This could be done through the creation of ad hoc forums that create safe spaces for companies (particularly civilian companies) and research institutions to share their expertise and views without fearing public relations concerns or commercial consequences.

Support an initiative on responsible AI for international peace and security

States that are members of the Alliance for Multilateralism could support the creation of an initiative on responsible AI for peace and security.[100] Such an initiative could form a bridge between other initiatives on responsible AI and the arms control discussion

[100] On the alliance see Alliance for Multilateralism, 'The Alliance for Multilateralism', [n.d.].

on AI. It could help to sensitize civilian actors beyond responsible AI initiatives to arms control objectives. In addition, it could provide an opportunity for states to receive useful input for their deliberations in the ongoing discussions on LAWS at the UN CCW.

Identify principles for responsible military use of AI

States should work to identify principles for the responsible military use of AI, and should codify them in official documents and, if possible, through collaborative multilateral processes. Such codified documents could provide a useful baseline for AI researchers to understand what constitutes legal, ethically acceptable and technically safe military end-use of AI technology and would allow for export controls to be adjusted accordingly.

Support education and training activities targeting actors in the AI sector

States should increase their financial support for initiatives that provide training on RRI to engineering students in the wider AI field. The project UNODA currently implements in the Asia–Pacific region could function as a model for such efforts in Europe and other regions.

Facilitate the participation of governmental experts with military and arms control expertise in responsible AI initiatives

States should support greater participation of governmental experts in multi-stakeholder events and initiatives related to RRI in AI. This should enable states to present government perspectives on international peace and security in discussions on what constitutes responsible behaviour in AI innovation and could result in cross-pollination between RRI processes and arms control processes.

About the authors

Dr Vincent Boulanin (France) is a Senior Researcher leading SIPRI's research on emerging military and security technologies. His focus is on issues related to development, use and control of autonomy in weapon systems and the military applications of artificial intelligence (AI). He regularly presents his and SIPRI's work, and engages with governments, United Nations bodies and other international organizations, research institutes and the media. Before joining SIPRI in 2014, he completed a doctorate in political science at the École des Hautes Études en Sciences Sociales, Paris. His recent publications include *Responsible Military Use of Artificial Intelligence: Can the European Union Lead the Way in Developing Best Practice* (co-authored, SIPRI: 2020), *Artificial Intelligence, Strategic Stability and Nuclear Risk* (co-authored, SIPRI: 2020), *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control* (co-authored, SIPRI/International Committee of the Red Cross: 2020).

Kolja Brockmann (Germany) is a Researcher in SIPRI's Dual-use and Arms Trade Control Programme. He first joined SIPRI as a European Union Non-Proliferation and Disarmament Consortium Intern and has been working at SIPRI since 2017. Prior to joining SIPRI, he was an intern with the German Federal Office for Economic Affairs and Export Control (BAFA) in Frankfurt. He graduated from King's College London with an MA in Non-Proliferation and International Security. At SIPRI, he conducts research in the fields of export control, non-proliferation and technology governance. His recent work has focused on controls on emerging technologies, in particular additive manufacturing (3D-printing) and intangible transfers of technology (ITT).

Luke Richards (United Kingdom) is a Research Assistant working on emerging military and security technologies. His current focus is on the responsible innovation and ethics of AI alongside broader technology governance issues. Prior to joining SIPRI, he worked at the International Institute for Strategic Studies (IISS) while finishing an MSc in Science and Technology Policy, writing his master's dissertation on the 'The Civil–Military Entanglement of Global Innovation'. He has previously worked on diverse projects, ranging from the development of a methodology to rank the cyber power of states through to the security implications of human enhancement.

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE
Signalistgatan 9
SE-169 72 Solna, Sweden
Telephone: +46 8 655 97 00
Email: sipri@sipri.org
Internet: www.sipri.org