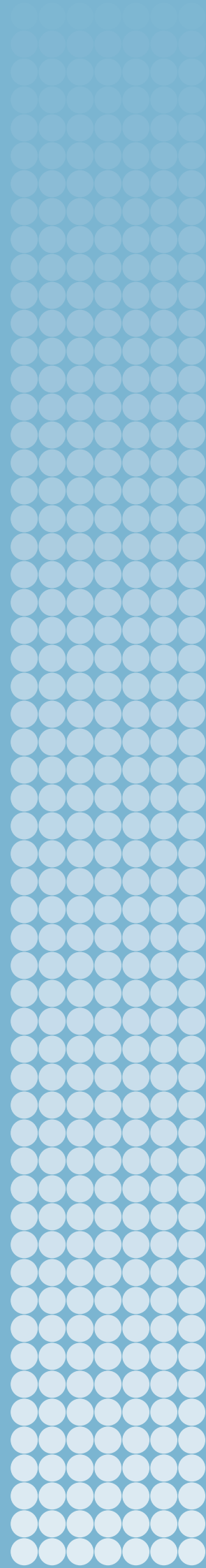


RESPONSIBLE ARTIFICIAL INTELLIGENCE RESEARCH AND INNOVATION FOR INTERNATIONAL PEACE AND SECURITY

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.

GOVERNING BOARD

Ambassador Jan Eliasson, Chair (Sweden)
Dr Vladimir Baranovsky (Russia)
Espen Barth Eide (Norway)
Jean-Marie Guéhenno (France)
Dr Radha Kumar (India)
Ambassador Ramtane Lamamra (Algeria)
Dr Patricia Lewis (Ireland/United Kingdom)
Dr Jessica Tuchman Mathews (United States)

DIRECTOR

Dan Smith (United Kingdom)



**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

<i>Acknowledgements</i>	v
<i>Executive summary</i>	vii
<i>Abbreviations</i>	ix
1. Introduction	1
Box 1.1. What is artificial intelligence?	2
2. Addressing the risks posed by the military use of AI	3
I. AI and international peace and security	3
Humanitarian and strategic risks	3
Risk vectors: Development, diffusion and use of AI technology	5
II. Addressing humanitarian and strategic risks using arms control	6
Arms control as a tool to govern the development, diffusion and military use of AI	7
Box 2.1. AI explainability and the black box problem	5
Figure 2.1. Foreseeable military applications of AI	4
Figure 2.2. Arms control as a process	6
3. Responsible research and innovation as a means to govern the development, diffusion and use of AI technology	11
I. RRI in the support of arms control on the military use of AI	11
RRI as an approach to technology governance	11
The advantages of RRI for technology governance and arms control	12
II. How would RRI in AI work in practice?	14
The knowledge needed	14
The means for implementing RRI	16
Identifying possible outcomes	17
4. Building on existing efforts to promote responsible research and innovation in AI	19
I. Building on existing responsible AI initiatives	19
Responsible AI initiatives	19
Challenges and opportunities	19
II. Building on export controls and compliance systems	25
Export control regulations and internal compliance programmes in academia, research institutes and the private sector	25
Challenges and opportunities	26
III. Conclusions on synergies between responsible AI initiatives and export control compliance	30
Box 4.1. Notable responsible AI initiatives	20
Figure 4.1. Frequently cited principles for responsible AI	22
5. Key findings and recommendations	31
I. Key findings	31
II. Recommendations	32
Companies, research institutes and universities	32
States and regional organizations	32
<i>About the authors</i>	34

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