CENTER FOR WAR STUDIES

sipri

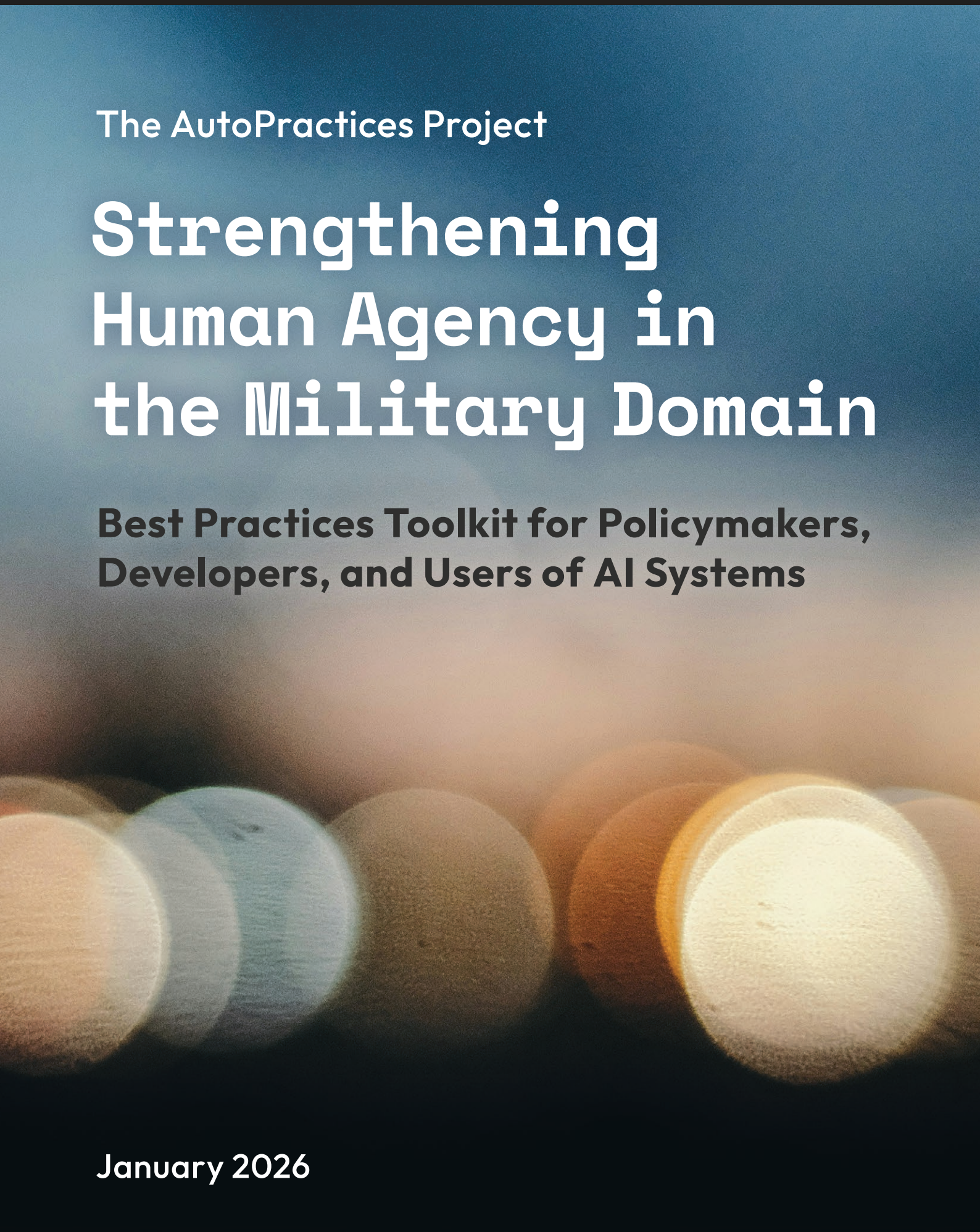
AutoNorms

SDU University of Southern Denmark

The AutoPractices Project

Strengthening Human Agency in the Military Domain

Best Practices Toolkit for Policymakers, Developers, and Users of AI Systems



January 2026

About the Center for War Studies

The Center for War Studies (CWS), established at the University of Southern Denmark (SDU) in 2012, brings together academics from political science, law, history, and cultural studies to contribute to major debates on the past, present, and future of war, as well as its impact on societies. We strive for interdisciplinary research that is relevant to policymakers and the society at large. We aim to contribute to ongoing debates on war and peace by illuminating their multiple dimensions. Through research excellence and societal relevance, we advance the understanding of the fundamental issue of war and peace. For more information about the CWS and its researchers, see https://www.sdu.dk/en/cws

Collaboration with SIPRI

The AutoPractices toolkit is published in collaboration with the Stockholm International Peace Research Institute (SIPRI). For more information about SIPRI, see https://www.sipri.org/

Funding

European Research Council
Established by the European Commission

AutoPractices is a Proof of Concept project funded by the European Research Council (grant number 101156237). It is based on the project "Weaponised Artificial Intelligence, Norms and Order" (AutoNorms), which has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 852123). For more information about AutoNorms and AutoPractices, see www.autonorms.eu

© Center for War Studies, January 2026

All rights reserved. No part of this publication may be reproduced, stored or transmitted, in any form or by any means, electronic and mechanical, without the prior written permission of the copyright holder or as explicitly permitted by law.

The AutoPractices team

Principal Investigator

Prof. Ingvild Bode, Professor of International Politics and Director of the Center for War Studies, SDU

Research Team

- Dr. Anna Nadibaidze, Postdoctoral Researcher, SDU
- Dr. Alexander Blanchard, Senior Researcher, SIPRI
- Shimona Mohan, in her personal capacity
- Ariel Conn, in her personal capacity
- Dr. Hendrik Huelss, Associate Professor, SDU
- Dr. Qiaochu Zhang, Max Weber Postdoctoral Fellow, European University Institute
- Dr. Guangyu Qiao-Franco, Assistant Professor, Radboud University Nijmegen
- Dr. Tom F.A. Watts, in his personal capacity

This practical toolkit was prepared by Anna Nadibaidze and Ingvild Bode, with input from the AutoPractices team and the stakeholders.

How to cite this publication:

The AutoPractices Project. (2026). *Strengthening Human Agency in the Military Domain: Best Practices Toolkit for Policymakers, Developers, and Users of AI Systems*. Odense: Center for War Studies, University of Southern Denmark. https://doi.org/10.5281/zenodo.17671715

Published by the Center for War Studies
University of Southern Denmark, Campusvej 55, Odense M 5230, Denmark

Cover photo by **insung yoon** | **@insungpandora**

Contents

Abbreviations

AI	artificial intelligence
AI DSS	AI-based decision-support systems
AWS	autonomous weapon systems
CCW	Convention on Certain Conventional Weapons (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)
CONOPS	Concepts of Operations
ERC	European Research Council
GGE on LAWS	Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems
ICRC	International Committee of the Red Cross
IEEE SA	The Institute of Electrical and Electronics Engineers Standards Association
IHL	international humanitarian law
LLM	large-language model
OECD	Organisation for Economic Co-operation and Development
REAIM	Responsible AI in the Military Domain summit
TEVV	Testing, evaluation, verification, and validation
UN	United Nations
UNGA	United Nations General Assembly
UNODA	United Nations Office for Disarmament Affairs
UUV	uncrewed underwater vehicle

Executive summary

Routine uses of artificial intelligence (AI) technologies affect military personnel's capacity to exercise human agency. The impacts of using AI on the exercise of human agency in the military domain should be addressed via bottom-up forms of governance. Such a process starts with the practices performed by stakeholders in technical, military, legal, and political fields.

This toolkit includes best practices which contribute to sustaining and strengthening the exercise of human agency in the development and use of military systems integrating AI technologies. These best practices concern personnel at various levels of military command, as well as those involved in the development, design, engineering, analysis, and other human-led fields across the lifecycle of AI systems.

The toolkit is the capstone output of the European Research Council (ERC)-funded AutoPractices project, which initiated a process of social innovation to govern AI technologies in military systems from the bottom up. This process involved the participation of 49 stakeholders representing diverse disciplines and all regions of the world. The stakeholders co-created the toolkit based on their experiences and fields of expertise.

The AutoPractices project takes as its starting point the assumption that the exercise of human agency in the development and deployment of AI systems should be sustained and strengthened across these systems' lifecycle.[1] The need for exercising a sufficient level of human agency is a recognised governance principle across global and regional initiatives on AI in the military domain, such as the United Nations Group of Governmental Experts on Emerging Technologies in the area of Lethal Autonomous Weapons Systems (UN GGE on LAWS) or the Responsible AI in the Military Domain (REAIM) summits.[2] Most initiatives in this space aim to set high-level legal and ethical principles on the human role in the use of force from the top down.

1 Ingvild Bode, Anna Nadibaidze, Tom F.A. Watts, and Qiaochu Zhang, "Ensuring the Exercise of Human Agency in AI-Based Military Systems: Concerns across the Lifecycle," *Ethics and Information Technology* 27, no. 50 (2025), https://doi.org/10.1007/s10676-025-09861-2.

2 Ingvild Bode, *Emerging Norms around Military Applications of AI: The Case of Human Control*, GC REAIM Expert Policy Note Series (The Hague Centre for Strategic Studies, 2025), https://hcss.nl/wp-content/uploads/2025/05/Bode-1.pdf; Ingvild Bode, *Human-Machine Interaction and Human Agency in the Military Domain*, Policy Brief No. 193 (Centre for International Governance Innovation, 2025), https://www.cigionline.org/publications/human-machine-interaction-and-human-agency-in-the-military-domain/; Anna Nadibaidze, Ingvild Bode, and Alexander Blanchard, "From Human Control to Human Agency in the Debate about Military Applications of Artificial Intelligence," preprint, *SSRN*, 2025, https://doi.org/10.2139/ssrn.5755743.

The bottom-up approach pursued by the AutoPractices project is complementary to these initiatives. It recognises that practices performed by various stakeholders in technical, military, and political spaces shape the implementation of high-level principles and thereby what key norms and guidelines mean at the practical level. At present, some of these practices appear to unintentionally lead to the acceptance of a compromised, diminished quality of human agency in the use of force.[3]

Yet, implementing best practices from the bottom up can also shape a positive version of human agency, ensuring a genuinely high quality of that agency rather than a nominal one. This toolkit identifies the practices that contribute to recontextualising the emerging norm of human agency into this positive version. AutoPractices' bottom-up approach to governing AI in the military domain is crucial for achieving the widely shared high-level principle of human agency over use-of-force decision-making.

This toolkit results from a co-creative process in which the stakeholders played an active role from start to finish. The AutoPractices team gathered practical insights from stakeholders via surveys, interviews, and workshops. Consequently, the practices compiled in the toolkit reflect those stakeholders' viewpoints. Throughout the project, stakeholders reviewed the two main AutoPractices outputs—the *Map of Practices* and the draft toolkit—and the AutoPractices team refined these documents based on stakeholders' feedback.[4]

This practical toolkit is intended for political decision-makers, designers, technical experts, and military personnel involved in the development and use of military systems integrating AI technologies. It may also be useful for researchers and scholars across various disciplines, civil society, and other stakeholders involved in the global debate on governing AI in the military domain.

The toolkit is particularly relevant for AI systems integrated into the complex process of use-of-force decision-making. However, it does not apply to any specific set of systems, whether autonomous weapon systems (AWS)[5] or AI-based decision-support systems (AI DSS).[6] Rather, the co-creation process involved a broad discussion, with the aim of "laying out a framework that can be translated by a lot of different people into practical implementation", as one stakeholder put it.

---

3 Ingvild Bode and Tom F.A. Watts, *Meaning-Less Human Control: Lessons from Air Defence Systems on Meaningful Human Control for the Debate on AWS* (Drone Wars UK & Center for War Studies, 2021), https://dronewars.net/2021/02/19/meaning-less-human-control-lessons-from-air-defence-systems-for-lethal-autonomous-weapons/; Ingvild Bode and Tom F.A. Watts, *Loitering Munitions and Unpredictability: Autonomy in Weapon Systems and Challenges to Human Control* (Center for War Studies & Royal Holloway Centre for International Security, 2023), https://www.autonorms.eu/loitering-munitions-and-unpredictability-autonomy-in-weapon-systems-and-challenges-to-human-control/.

4 See The AutoPractices Project, *Map of Practices* (Center for War Studies, 2025), https://www.autonorms.eu/governing-ai-technologies-in-the-military-domain-from-the-bottom-up-map-of-practices/.

5 There is no universally agreed definition of (lethal) autonomous weapon systems. In this document, AWS are defined as weapon systems that select and engage targets without human intervention. See International Committee of the Red Cross, "ICRC Position on Autonomous Weapon Systems," May 12, 2021, https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems.

6 There is no universally agreed definition of AI DSS. In this document, AI DSS are defined as tools integrating AI technologies employed for various tasks part of military targeting decision-making. For a more detailed discussion of AI DSS and their operation, see Anna Nadibaidze, Ingvild Bode, and Qiaochu Zhang, *AI in Military Decision Support Systems: A Review of Developments and Debates* (Center for War Studies, 2024), https://www.autonorms.eu/ai-in-military-decision-support-systems-a-review-of-developments-and-debates/; Alexander Blanchard and Laura Bruun, *Autonomous Weapon Systems and AI-Enabled Decision Support Systems in Military Targeting: A Comparison and Recommended Policy Responses* (Stockholm International Peace Research Institute, 2025), https://doi.org/10.55163/YQBY3151; Arthur Holland Michel, *Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making* (International Committee of the Red Cross, 2024), https://shop.icrc.org/decisions-decisions-decisions-computation-and-artificial-intelligence-in-military-decision-making-pdf-en.html. On AI and intelligence analysis, see Alexander Blanchard and Mariarosaria Taddeo, "The Ethics of Artificial Intelligence for Intelligence Analysis: a Review of the Key Challenges with Recommendations," *Digital Society* 2, no. 12 (2023), https://doi.org/10.1007/s44206-023-00036-4.

Summary of recommendations and take-aways

The various practices along the lifecycle that the AutoPractices co-creation process has revealed are the recommendations we put forward to sustain and strengthen the exercise of human agency. Based on these practices, we highlight four key take-aways for policymakers, developers, and users of AI-based systems in the military domain:

A Establish boundaries at the early stages of the AI lifecycle

Determining specific objectives to the development and use of AI systems, including limitations and boundaries to the applications of these systems, early in the lifecycle, strengthens the exercise of human agency and humans' ability to make meaningful choices down the line.

B Implement feedback and communication mechanisms across the AI lifecycle

Constant feedback and communication mechanisms between actors involved at various lifecycle stages contribute to enabling human understanding of AI systems and their uses within specific contexts. This would enhance transparency and thereby human accountability, as well as prevent undue AI influence.

C Prioritise education and training across the AI lifecycle

The continuous education and training of humans across lifecycle stages (policymakers, developers, testers, operators, and users, among others) play a crucial role in raising humans' AI literacy and awareness about dynamics of human-AI interaction. This mitigates some of the risks arising from these dynamics in operating contexts and contributes to strengthening the exercise of human agency within those contexts.

D Adopt strategies and steps to implement the practices listed in this toolkit

Developing and adopting their own, concrete approaches to operationalising the practices presented in this toolkit would allow policymakers to go beyond high-level principles on responsible and ethical AI, while accounting for specific types of systems, conditions, and contexts.

Box 1 Key terms

Human agency For the purposes of the AutoPractices project, the exercise of human agency in the context of interacting with AI systems is defined as: *The capacity to 1) understand and reasonably foresee a system's functions and effects in a relevant context; and 2) to deliberate and decide upon suitable actions in a timely manner; and 3) to act in a way that can impact the use of the system.*

AI For the purpose of this toolkit, we follow the definition put forward by the OECD, where an AI system is defined as "a machine-based system, that for explicit or implicit objectives infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions".[7] Practices outlined in this toolkit are based on stakeholders' knowledge of and expertise about currently available technological developments in the field of AI. Future developments of AI technologies, such as in agentic AI,[8] are likely to have further implications for human agency, and will likely require the adaptation of some practices.

The AI lifecycle The AutoPractices project takes as a foundation the model of the AI lifecycle proposed by the IEEE Standards Association Research Group on Issues of Autonomy and AI in Defense Systems.[9] The framework presents a granular way of thinking about the lifecycle with multiple points of human involvement and opportunities to exercise agency. In this model, the lifecycle of a military system integrating AI technologies includes the following 9 stages, which do not necessarily occur in a linear way.

1. Before system development
2. Research and development
3. Procurement and acquisition
4. Testing, evaluation, verification, and validation (TEVV)
5. Considering the human: Education, training, and human-system integration
6. Political and strategic considerations
7. Operational-level command and control
8. Tactical employment
9. Review, reuse, and/or retire.

Any reference to 'lifecycle' throughout the document refers to this model of the AI lifecycle, rather than procurement lifecycles, machine learning lifecycles, or systems engineering lifecycles. For more information about the key terms used in AutoPractices outputs, see Appendix C.

7 OECD Committee on Digital Economy Policy, *Explanatory Memorandum on the Updated OECD Definition of an AI System*, OECD Artificial Intelligence Papers No.8 (OECD, 2024), https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51/623da898-en.pdf.

8 AI agents can be defined as "AI systems capable of acting without direct human guidance". See Vincent Boulanin, Alexander Blanchard, and Diego Lopes da Silva, "Before It's Too Late: Why a World of Interacting AI Agents Demands New Safeguards," *Stockholm International Peace Research Institute*, October 1, 2025, https://www.sipri.org/commentary/essay/2025/its-too-late-why-world-interacting-ai-agents-demands-new-safeguards.

9 IEEE SA Research Group on Issues of Autonomy and AI in Defense Systems, *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications* (IEEE SA, 2024), https://ieeexplore.ieee.org/document/10707139.

Acknowledgements

We are grateful to all 49 stakeholders who participated in the AutoPractices project, whether by filling out the survey, participating in an interview, or joining the workshops. All stakeholders have been invited to contribute in their personal capacity.

This toolkit is meant to capture the co-creation process as accurately as possible. While it aims to be representative of stakeholders' contributions, the toolkit does not necessarily reflect the views of all stakeholders, the AutoPractices team, or the organisations they are affiliated with.

The following stakeholders gave explicit consent to be acknowledged by name for their participation in the project (in alphabetical order):

Zena Assaad, Rupert Barrett-Taylor, Ishmael Bhila, Joanna Bryson, Dylan Cawthorne, Neil Davison, Jessica Dorsey, Toni Erskine, Denise Garcia, Netta Goussac, Manoj Harjani, Adam Hepworth, Elisabeth Hoffberger-Pippan, Arthur Holland Michel, Keketso Kgomosotho, Joanne Kirkham, Klaudia Klonowska, Anna Knack, Atay Kozlovski, Jonathan Kwik, Dustin Lewis, Branka Marijan, Hugo May, Ann-Katrien Oimann, Philippe Prince Tritto, Ioana Puscas, Haotian Qi, Gregory Reichberg, Thea Riebe, Laure de Roucy-Rochegonde, Camilo Serna, Jack Shanahan, Peter Spayne, Chirayu Thakkar, Jimena Sofía Viveros Álvarez, Nathan Wood, Ling Zhang.

Toolkit of best practices to strengthen human agency across the AI lifecycle

This section contains sets of best practices identified by AutoPractices stakeholders as tools that contribute to sustaining and strengthening human agency in the context of developing and using AI systems in the military domain.

The toolkit spells out the practices in bold and thereafter offers more details on what the practice entails and how it relates to the three elements of the human agency definition (see Box 1), if such details were provided by the stakeholders. The toolkit includes direct quotes from stakeholders and cites stakeholders' publications when the latter are relevant to the practices and their implementation, and only when stakeholders gave explicit consent to inlude these quotes and publications.

Contributions of the toolkit

This toolkit's main contribution is to highlight practical ways to sustain and strengthen the exercise of human agency by following a bottom-up approach to governance of AI in the military domain.

Many practices listed in this toolkit are already implemented by defence organisations in relation to other military systems. Yet, it is important to highlight that these practices also apply to systems integrating AI. Some of these practices also require renewed attention for AI systems. They should not be considered in isolation. As AI systems are often employed as part of a network, in combination with other types of systems, it is crucial to see existing and new practices as being interrelated.

The use of AI systems that are increasingly integrated into existing sets of practices and networks could have cascading effects and potentially involve new types of risks which need to be mitigated. Part of the AutoPractices exercise is therefore to clarify that many "old" practices continue to hold and should be performed in relation to AI systems, especially as the use of these systems involve particular challenges to the exercise of human agency (see Appendix A).

These best practices are not to be conflated with legal obligations set by international humanitarian law (IHL) or international conventions that states are party to, for example, Article 36 of the 1977 Additional Protocol I to the 1949 Geneva Conventions.

Structure of the toolkit

The toolkit is structured along the 9 stages of the AI lifecycle (as defined in Box 1), with practices associated to each stage. Practices for stages 7 (operational-level command and control) and 8 (tactical employment) are combined into one section given the natural repetition and overlap in the practices associated with these stages.

This structure of the toolkit allows readers to focus on lifecycle stages that they are most involved in. It is intended to ensure the toolkit is as easy to use as possible. After covering the 9 stages, the toolkit includes sets of practices that cut across different stages, highlighting that the lifecycle is not always a linear process.[10]

At each stage, the toolkit mentions groups of human actors who are likely to be substantially involved. Implementing this toolkit's best practices will differ across actors' organisational, military, political, and societal cultures, what these cultures prioritise, as well as the level of integration between the military and the private sector, especially the technology industry. We also recognise that capacity limitations affect the implementation process of many practices, as actors around the world have different types of resources (financial, human, technical, or other) available for this work. Mitigating the multiple challenges in the implementation of these practices involves capacity building measures which are outside the scope of this toolkit.

Table 1 illustrates how different practices identified by stakeholders and listed according to the 9 AI lifecycle stages contribute to the three main elements of human agency, which in the context of AutoPractices is defined as *the capacity to*

- *understand and reasonably foresee a system's functions and effects in a relevant context (element 1); and*
- *to deliberate and decide upon suitable actions in a timely manner (element 2); and*
- *to act in a way that can impact the use of the system (element 3).*

This definition it is not only directed to users of AI systems, but to groups of different humans involved across the lifecycle.[11]

The definition, as the toolkit and the process that led to it, follow the assumption that human agency is exercised with intentions based on the principle of humanity as a hallmark of IHL, rather than with the intentions to commit harm or violations of applicable domestic and international law.

10 As the IEEE SA Research Group on Issues of Autonomy and AI in Defense Systems notes, while "some of the nine lifecycle stages are necessarily sequential…many [stages] may have significant overlap, and may start or end at around the same time". See IEEE SA, *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*, 8.

11 Stakeholders remarked that the definition appears to be focused on the time of use. Interestingly, although some stakeholders had this impression, most of their responses across the different forms of data collection focused on early lifecycle stages. This underlines the importance of the early lifecycle stages for sustaining and strengthening the exercise of human agency.

Table 1 Overview of practices and their contributions to elements of human agency

Practice	E 1	E 2	E 3
Stage 1: Before development			
A Set clear objectives and boundaries for the development of the systems.	X		
B Clearly conceptualise the anticipated circumstances and environments of use to the extent possible, and, if needed, delimit the use cases.	X	X	X
C Determine whether the systems should be used in the planned contexts and whether the routine uses of the systems in these contexts involve risks of atrophying human capabilities.	X		
D Design a comprehensive risk assessment framework and make (political) choices regarding risk acceptance and risk thresholds.	X	X	X
E Classify the role of the systems in the use of force.	X	X	X
F Design mechanisms of human intervention that are appropriate to the risks involved and the role of the systems in the use of force.		X	X
G Set rigorous criteria for assisting the identification of risks and errors that might arise.	X		
H Engage in joint efforts with other states or organisations to identify best practices and benchmarks for AI systems to meet objectives while assessing risks.	X		
I Choose transparent testing procedures that embed decision traceability within them.	X		
J Foster organisational cultures that promote feedback and exchange.	X	X	X
K Decide on trade-offs related to the interface design and how it presents information to humans involved.	X	X	
Stage 2: Research and development			
A Record limitations or biases within the training data via robust documentation and ensure that documentation is read and understood.	X	X	
B Implement the risk assessment frameworks and mechanisms for human intervention based on the risk thresholds set previously.		X	X
C Design interfaces that reflect previously set limitations and objectives and present the information appropriately.	X	X	X
D Design interfaces that provide ways of identifying system errors and knowing when systems are not performing as expected.	X	X	X
E Integrate monitoring mechanisms for users to identify when systems are not working according to planned objectives and considerations.	X	X	
F Involve users in the design of the interface.		X	X
G Design adaptive interfaces that support education measures.		X	X
Stage 3: Procurement and acquisition			
A Implement monitoring mechanisms to ensure that the systems acquired function as intended and in line with the objectives set previously.	X	X	
B Implement the risk assessment frameworks and mechanisms for human intervention based on the risk thresholds.		X	X
C Involve users to identify the appropriate level of human intervention from an operational perspective.		X	X
Stage 4: TEVV			
A Test the interface.	X	X	X
B Implement the risk assessment frameworks and mechanisms for human intervention based on the risk thresholds.		X	X
C Use precise metrics for evaluation in various scenarios, not merely follow a check list approach.	X	X	X
D Test the systems to verify that the intended users have access to all the required information and are empowered to make legally compliant decisions and exercise ethical judgement based on this information.		X	
E Continuously test with independent audits and without over-trusting commercial vendors' claims about AI systems.	X	X	X

Practice	E 1	E 2	E 3
Stage 4: TEVV (cont)			
F Test and update AI systems with decision traceability.	X	X	
G Use appropriate datasets.	X		
H Involve actors beyond technical specialists, including lawyers, ethicists, and operators.	X	X	X
Stage 5: Considering the human			
A Train humans involved across the lifecycle on the objectives of AI systems set at earlier stages.	X	X	
B Train operators to use the interface in a way that complies with objectives and limitations set previously.	X		
C Continuously train on cross-checking and critically assessing the outputs of AI systems.		X	
D Continuously train on the dynamics of human-AI interaction.	X	X	
E Ensure a balance between sufficient understanding and appropriate calibration of trust in AI systems' outputs.	X	X	
F Ensure ways for users to understand when systems are not working according to planned objectives and requirements.	X	X	
G Adapt training measures dynamically when AI systems develop or are updated.	X		
Stage 6: Political and strategic considerations			
A Raise awareness among political decision-makers about the benefits and limitations of AI systems without overhyping or understating AI technologies.	X		
B Incorporate concerns about the escalation risks of military applications of AI and misleading narratives amid geopolitical rivalry.	X	X	
C Assess the conduct of practices within adversarial contexts.		X	
Stage 7 & Stage 8: Operational-level command and control & Tactical employment			
A Ensure the deployment of AI systems is limited to anticipated circumstances and environments, to the extent possible.	X	X	X
B Follow safeguards and standards designed for users to understand if the systems do not work as planned.		X	X
C Implement the risk assessment frameworks and mechanisms for human intervention based on the risk thresholds.		X	X
D Continuously evaluate the development of the systems during use and relaying feedback to TEVV.	X	X	
E Log the targeting process and create appropriate levels of documentation.	X		
F Collect information during use, as in the case of 'black boxes' used in the civilian aviation context.	X		
Stage 9: Review, reuse, and/or retire			
A Implement reviews to evaluate whether previously established political, strategic, legal, ethical, and technical considerations have been met.	X	X	X
B Process data collected during use, compare the outcomes of the systems' use to risk assessment frameworks, and document whether the systems can be reused (in the case of hardware), actively used, paused, discontinued, or need to be updated.	X		
C Communicate the findings to decision-makers and developers to support feedback loops.	X	X	X
D Actively review and retest following system updates.	X		
Cross-cutting practices (across the lifecycle)			
A Train on AI systems and human-AI interaction.	X	X	X
B Ensure communication and organisational learning.	X	X	X
C Apply comprehensive risk assessment and mitigation.	X	X	X
D Strengthen cybersecurity and reliability of AI systems.	X	X	X
E Implement additional safeguards for online learning.	X	X	X

The elements are also included in parentheses at the end of each practice's description (1,2,3).

Stage 1: Before system development

Practices at this stage will mainly concern political and strategic decision-makers.

A Set clear objectives and boundaries for the development of the systems. Setting clear and concrete goals for the development of AI systems within the planned contexts and environments of use improves humans' ability to consider how the AI systems, and different configurations of human–AI interaction, relate to the contexts as well as broader political, strategic, ethical, and societal considerations, and legal obligations. The explicit objectives defined in guiding documents such as Use Cases and Concepts of Operations (CONOPS) are essential for preventing faulty assumptions and establishing guardrails against undesirable applications of AI. This includes cases when AI capabilities are to be integrated into existing military systems or processes. Identifying uses and applications where AI technologies are unsuitable is a key part of this practice. This practice also involves setting boundaries on systems' capabilities in particular operational environments. Clear delimitations contribute to operators' ability to predict the functioning of the systems in the relevant contexts or theatres of operation. (1)

B Clearly conceptualise the anticipated circumstances and environments of use to the extent possible, and, if needed, delimit the use cases. Identifying the anticipated domains of use (e.g., land, air, subsea, space) and related circumstances (to the extent possible) is crucial to strengthening humans' understanding of the systems and how they might function within those environments. Limitations to uses of AI systems can be geographical and temporal in nature or refer to types of targets or target profiles, among others. This practice involves defining expected and allowed uses in advance.[12] This includes developing basic guidance and protocols for users based on these limitations, for instance as part of Use Cases and CONOPS, as well as developing a guide for roles of humans across the lifecycle.[13] (1, 2, 3)

C Determine whether the systems should be used in the planned contexts and whether the routine uses of the systems in these contexts involve risks of atrophying human capabilities. This includes identifying use cases within those contexts where there is a higher risk of human skills being affected, as well as ways of mitigating these challenges. Involving experts across different disciplines (across the technical and social sciences) in this discussion provides a comprehensive perspective on the variety of potential risks related to configurations of human–AI interaction. (1)

D Design a comprehensive risk assessment framework and make (political) choices regarding risk acceptance and risk thresholds.[14] Such a framework can reflect the risk assessment procedures that are already part of operational frameworks, planning tools, rules of engagement, and targeting doctrines. It may also involve updating existing frameworks, Use Cases, and CONOPS with specific guidelines as part of "standardised operating procedures and fragmentary orders", as well as checklists and flowcharts that integrate

12 On system card templates and use-case libraries for defence organisations, see Anna Knack et al., *Defence AI Assurance: Identifying Promising Practice and a System Card Template for Defence* (The Alan Turing Institute, 2025), https://www.turing.ac.uk/sites/default/files/2025-09/defence_ai_assurance.pdf.

13 On the 'Intended Operational Environment' see Jonathan Kwik, *Lawfully Using Autonomous Weapon Technologies* (T.M.C. Asser Press, 2024), 71–72 and 385; on operational considerations see Jack Shanahan, *The Need for A Lifecycle Approach to AI-Enhanced Autonomy*, Defense Paper Series (Special Competitive Studies Project, 2025), https://www.scsp.ai/wp-content/uploads/2025/09/Shanahan_Defense-Paper-Series.pdf.

14 See the ongoing activity 'risk assessments' in IEEE SA, *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*.

directives for various levels of risks.[15] Considering the differences and distinctive features of types of conflicts when assessing the risks is an important part of designing risk assessment frameworks and the technical, policy, and procedural practices to be adopted depending on the level of risk.[16] Involving multidisciplinary stakeholders in the design or update of risk assessment frameworks allows considering risks affecting human capabilities in relation to all humans involved—not only the prospective users of a system.[17] Risk assessment frameworks present a productive way of ensuring the integration of a throughline of ethical, legal, security, and other concerns from design to deployment.[18] (1, 2, 3)

E **Classify the role of the systems in the use of force.** This practice involves understanding the functions which the AI-based systems are tasked with performing in use-of-force decision-making. For example, in the case of AI DSS, it could be analysing data, recognising patterns within the data, or recommending courses of action. This classification's value lies in strengthening humans' understanding of the systems' role(s) in the intended contexts of use, as well as informing risk assessment frameworks and therefore responses that are fitting to address the risks and the context. Considering unintended impacts of AI systems on humans' decision-making processes across the chain of command (e.g., an error being carried down the chain) is valuable, as it strengthens the understanding of potential effects of this decision-making and potential responses. (1, 2, 3)

F **Design mechanisms of human intervention that are appropriate to the risks involved and the role of the systems in the use of force.** Establishing risk thresholds lets the actors involved tailor implementation mechanisms to require varying degrees of human intervention, thereby mitigating risks according to each threshold's severity. Human intervention in the development and use of AI systems can be conceptualised as a spectrum, where a higher risk threshold demands a higher level of human intervention. Practices establishing mechanisms to ensure these levels of human intervention contribute to strengthening the exercise of human agency at the use stages (6–9), especially if they provide guidance on conducting assessments, engaging in critical reflections, questioning the outputs of AI systems, and eventually intervening. Practices to implement mechanisms of human intervention include cross-checking the outputs of AI systems, limitations on the overall speed of decision-making, and ways for humans to override or potentially deactivate AI systems. (2, 3)

15 Kwik, *Lawfully Using Autonomous Weapon Technologies*, 389–96; on system card templates see Knack et al., *Defence AI Assurance: Identifying Promising Practice and a System Card Template for Defence*; on rules of engagement and targeting doctrines see Jonathan Kwik, Marten Zwanenburg, Cheya Took, and Joep Aarts, "Controlling Military Artificial Intelligence: Harnessing Rules of Engagement and Military Directives," ASSER Policy Brief No. 2025-01, *SSRN*, 2025, https://doi.org/10.2139/ssrn.5132731.

16 Jack Shanahan, "Symposium on Military AI and the Law of Armed Conflict: A Risk Framework for AI-Enabled Military Systems," *Opinio Juris*, April 1, 2024, https://opiniojuris.org/2024/04/01/symposium-on-military-ai-and-the-law-of-armed-conflict-a-risk-framework-for-ai-enabled-military-systems/.

17 Stakeholders recommend a comprehensive risk assessment because there are a variety of risks to the exercise of human agency beyond those that are commonly cited in the debate on AI technologies, such as biases or malfunctions. For instance, an output generated by an AI-based system may be technically correct, but not accurate or appropriate in the particular context of use.

18 See Zena Assaad, "A Risk-Based Trust Framework for Assuring the Humans in Human-Machine Teaming," *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, ACM, 2024, 1–9, https://doi.org/10.1145/3686038.3686045; Alexander Blanchard, Claudio Novelli, Luciano Floridi, and Mariarosaria Taddeo, "A Risk-Based Regulatory Approach To Autonomous Weapon Systems," *Digital Society* 4, no. 1 (2025), https://doi.org/10.1007/s44206-025-00181-y; Ioana Puscas, *AI and International Security: Understanding the Risks and Paving the Path for Confidence-Building Measures* (United Nations Institute for Disarmament Research, 2023), https://unidir.org/publication/ai-and-international-security-understanding-the-risks-and-paving-the-path-for-confidence-building-measures/.

G **Set rigorous criteria for assisting the identification of risks and errors that might arise.** This practice covers considering conditions about how long the systems could be used without needing to be reviewed, updated, or re-tested. Detailed guidance for identifying risks strengthens users' understanding of the system and their awareness surrounding risks and potential malfunctions. (1)

H **Engage in joint efforts with other states or organisations to identify best practices and benchmarks for AI systems to meet objectives while assessing risks**. Sharing lessons learned can advance comprehensive thinking and practices around an emergent risk assessment framework, as well as strengthen interoperability. (1)

I **Choose transparent testing procedures that embed decision traceability within them.** This includes deciding on the appropriate level of documentation to record the testing processes and ensure the transparency of these processes. It might involve creating or using organisational frameworks to conduct audits. Making these decisions at stage 1 provides clearer guidance for all humans involved across the lifecycle, strengthening their knowledge of the systems. (1)

J **Foster organisational cultures that promote feedback and exchange.** This practice includes discussing AI uses in various contexts and sharing experiences of interacting with AI systems. It also involves raising humans' awareness about both positive and adverse effects of human-AI interaction within the organisation's respective culture and structure of learning. (1, 2, 3)

K **Decide on trade-offs related to the interface design and how it presents information to humans involved.** This practice recognises that small design choices around how the interface functions, how information is displayed visually and how much information is displayed can have potential "domino effects" for the exercise of human agency. Making such design choices prior to development involves navigating trade-offs between providing sufficient information to sustain human agency and preventing cognitive overload of the humans involved. The interface design plays a role in strengthening humans', including analysts, commanders, and operators, capacity to reasonably foresee a system's functions and deliberate upon suitable courses of action. Interface design choices will likely need to be revisited throughout the latter stages of the lifecycle, especially in relation to feedback loops. (1, 2)

Stage 2: Research and development

Practices at this stage will mainly concern designers, researchers, and developers.

A **Record limitations or biases within the training data via robust documentation and ensure that documentation is read and understood.** This includes specifying whether the training data cover a particular geographical context and other potential biases related to gender, ethnicity, or age. Such limitations would inform decisions about where the systems should and should not be used. For instance, when an AI system is trained mainly on data from a single region, it should not be assumed that it will function as intended in other parts of the world. Such documentation could also cover where and how the data were sourced and whether they are susceptible to poisoning, in other words, causing significant imbalances in the data. Informing humans across the lifecycle about these limitations enables them to better foresee AI systems' impacts and deliberate on appropriate responses. This practice also includes various types of procedural thinking that goes into the selection and curation of training data. Key questions that could be addressed and documented at the organisational level include: Who selects training data? Who does the evaluation of the training data? Who labels training data? What are metrics for fairness that are used for assessment of bias? Data-related practices are particularly important in the military domain as a high-stakes, safety-critical context.[19] (1, 2)

B **Implement the risk assessment frameworks and mechanisms for human intervention based on the risk thresholds set previously.** Design choices will depend on the safeguards that are suitable for various levels of risks and thresholds. During the design phase, this involves incorporating mechanisms that let humans step in or enable AI systems to present several courses of action. Such approaches provide humans with the opportunity to discuss the options available and take timely actions based on those deliberations, if needed. (2, 3)

C **Design interfaces that reflect previously set limitations and objectives and present the information appropriately.**[20] A balance between sufficient information and cognitive overloading is needed for humans to understand and reasonably foresee AI systems' effects. Interfaces that offer humans multiple options or courses of action strengthen humans' ability to deliberate upon different actions and eventually choose the most appropriate action in relation to the context and environment of use. Priming or nudging techniques, meanwhile, act in the opposite way and would undermine humans' capacity to consider and decide upon appropriate responses. In the case of AI systems used to support the targeting process, this practice can include interfaces displaying "what constitutes the digital signature that creates a target", that is, the triangulated data points that led to identifying or selecting a particular target. (1, 2, 3)

D **Design interfaces that provide ways of identifying system errors and knowing when systems are not performing as expected.** This practice involves creating interfaces that let humans, especially operators and users, anticipate a system's effects in specific contexts and provide key information so they can consider and decide on appropriate responses. An interface design that supports the formation of an accurate mental model of the systems contributes to this practice. A balanced trust approach, with preselection to

19 Arthur Holland Michel, *Known Unknowns: Data Issues and Military Autonomous Systems* (United Nations Institute for Disarmament Research, 2021), https://unidir.org/known-unknowns.

20 Ioana Puscas, *Human-Machine Interfaces in Autonomous Weapon Systems* (United Nations Institute for Disarmament Research, 2022), https://unidir.org/publication/human-machine-interfaces-in-autonomous-weapon-systems/.

avoid information overload but not simplification or limitation of choices or courses of action presented to humans, is particularly valuable to all three elements of exercising human agency. (1, 2, 3)

E **Integrate monitoring mechanisms for users to identify when systems are not working according to planned objectives and considerations.** This practice includes designing AI-specific standards and safeguards for users to understand that the systems do not work as planned within the relevant context. Such safeguards and mechanisms provide guidance for humans, especially users and operators, thereby strengthening their ability to act in case of malfunctions or other risks arising from the use of AI systems. (1, 2)

F **Involve users in the design of the interface.** This practice aims to ensure that key design choices fulfil their intended functions and advance usability. Involving users allows designers and developers to identify the appropriate level of human intervention from an operational perspective (see also stage 3, practice C). (2, 3)

G **Design adaptive interfaces that support education measures**. This practice underlines the need for an adaptive design of AI systems that accounts for human cognitive loads. Sustained operations of AI systems in various military contexts raise concerns about decision fatigue thresholds and cognitive overload, strengthening the risks of automation bias in human-AI interaction. This could be addressed by modelling user stress and fatigue to support operator resilience at the points of both system and interface design. The practice recognises that concerns surrounding human cognitive (over)load become more pronounced when users simultaneously "work with multiple systems in multiple operational contexts". The interface design plays a critical role in supporting training and education measures, where humans including operators have the opportunity to strengthen their capacity to deliberate upon courses of action and act upon these deliberations in various environments of use (see stage 5, practice B). (2, 3)

Stage 3: Procurement and acquisition

Practices at this stage will mainly concern officials responsible for procurement in military organisations.

A **Implement monitoring mechanisms to ensure that the systems acquired function as intended and in line with the objectives set previously.** This practice entails ensuring that the objectives of developing and using AI systems set at earlier stages are not overtaken by vendors' and suppliers' priorities.[21] When resources and infrastructure allow, developing and maintaining a diverse AI ecosystem contributes to preventing being 'locked into' working with certain companies and suppliers who might have diverging priorities.[22] The practice includes identifying potential ways of how the systems procured might not comply with previously defined objectives and boundaries, as well as communicating these findings to relevant humans to bolster their understanding of how systems might behave in certain contexts and how to respond accordingly. (1, 2)

B **Implement the risk assessment frameworks and mechanisms for human intervention based on the risk thresholds.** This practice covers following a detailed risk mitigation approach, based on the risk assessments frameworks at earlier stages, and determining whether the systems acquired match

21 Netta Goussac, "Responsible Behaviour in Military AI Starts with Responsible Procurement," *Stockholm International Peace Research Institute*, October 16, 2025, https://www.sipri.org/commentary/essay/2025/military-ai-responsible-procurement.

22 Anna Knack, Nandita Balakrishnan, and Timothy Clancy, *Applying AI to Strategic Warning* (Centre for Emerging Technology and Security, 2025), 35, https://cetas.turing.ac.uk/publications/applying-ai-strategic-warning.

the planned objectives and context of use, especially when procured from outside actors. Depending on these assessments, mechanisms for humans to intervene might be needed to ensure that humans have the option to deliberate upon actions and act upon their decisions. This practice also involves considering the risks of relying on software supplied by private vendors. Nurturing a diverse AI ecosystem can contribute to reducing dependence on particular suppliers, and therefore reliance on vendors' preferences which might involve risks or not comply with the objectives and intended contexts of use. (2, 3)

C **Involve users to identify the appropriate level of human intervention from an operational perspective**. Users' input at this stage assists procurement officials with understanding the functions of AI systems, as well as deciding on the appropriate levels of human intervention and mechanisms which give users the opportunity to deliberate upon responses and act upon them. This practice recognises a balance between AI systems performing functions to address some of the cognitive load of users in analysing and combining data and ensuring that users are "in the right mental state to process information and to make a decision that they feel responsible for". Translating this balance into concrete policies based on intended users' input means that the procurement stage complies with objectives for and limitations to the use of AI systems set at earlier stages. (2, 3)

Stage 4: Testing, evaluation, verification, and validation (TEVV)

A **Test the interface**. This practice involves users in the testing of the interface. It includes identifying any limitations of the systems and documenting them. This extends to monitoring what the interface means for users' decision-making, including their ability to comply with legal, ethical, and operational considerations. Testing the interface allows TEVV personnel to assess how various configurations of human-AI interaction may impact humans' understanding of the systems and their ability to deliberate and act upon the information provided via the interface, including whether they risk atrophying human capabilities. TEVV and interface testing should not only focus on "whether the system works technically, but on how well human users can understand and interact with it". (1, 2, 3)

B **Implement the risk assessment frameworks and mechanisms for human intervention based on the risk thresholds**. This practice features continuously testing mechanisms for human intervention which were previously designed, in conditions that match the foreseen contexts of use to the extent possible, as well as potential unforeseen edge cases. This practice allows testers to evaluate how AI systems' performances relate to risk thresholds set previously, and whether the mechanisms for human intervention are appropriate for the intended contexts and environments of operation. (2, 3)

C **Use precise metrics for evaluation in various scenarios, not merely follow a check list approach.** This practice aligns with long-established TEVV practices of interrogating and testing systems but simultaneously considers specific challenges raised by AI systems. Concrete benchmarks for operational effectiveness, legal and ethical considerations, and safety protocols assist with strengthening humans' understanding of systems' functioning. This includes evaluating response mechanisms in the intended operational contexts and in other scenarios which might involve unexpected outputs. At the same time,

Practices at this stage will mainly concern TEVV experts, a broad category of actors that includes systems engineers, computer scientists, but also legal and ethical advisers. While TEVV encompass a specific stage of the lifecycle, it is also a set of practices that are ongoing through the lifecycle in a non-linear way (TEVV, monitoring, hardware system or software updates and interoperability, maintenance).

this practice recognises that there is a risk of personnel merely following a checklist of legal and ethical frameworks that are "instrumentalised in support of delivery". Comprehensive TEVV may allow fundamentally altering AI systems if evaluation does not meet the applicable benchmarks—including the choice of not deploying the systems. (1, 2, 3)

D **Test the systems to verify that the intended users have access to all the required information and are empowered to make legally compliant decisions and exercise ethical judgement based on this information**. This set of practices includes verifying systems' impacts upon users' abilities to comply with IHL principles and ethical considerations that are applicable,[23] not only in intended contexts of use but also in other scenarios. Such testing practices contribute to strengthening the users' ability to deliberate upon actions in relation to edge cases and unexpected or ambiguous outputs. They enhance the capacity for military personnel to use AI systems in a legally and ethically compliant way, while increasing operational effectiveness.[24] It is important to consider that technical performance does not equal legal compliance or operational effectiveness. (2)

E **Continuously test with independent audits and without over-trusting commercial vendors' claims about AI systems.** Conducting TEVV by independent experts aims at ensuring that AI systems are not deployed before having reached the required level of maturity, in accordance with objectives and risk assessment frameworks set earlier. For governments seeking to procure AI systems for military use, this practice includes conducting additional TEVV rather than solely relying on existing documentation and test results, including from allies. Existing test results may be a consideration, but not the only consideration. Continuous independent testing involves monitoring whether users of AI systems understand their functioning, limitations, and the information AI systems display in a way that does not confuse or overwhelm the human users. (1, 2, 3)

F **Test and update AI systems with decision traceability**. Regular reviews guided by metrics and technical audits include verifying for model and software updates, active review and retest procedures following selected types of updates, and maintenance processes that integrate feedback reports from operators. Such reviews are key to address ongoing changes in both the AI systems and the intended operational contexts which may affect the exercise of human agency. Appropriate levels of documentation in relation to malfunctions or incidents contribute to transparency and tracing decisions made at the TEVV stage. (1, 2)

G **Use appropriate datasets.** This practice entails using datasets that are appropriate to the intended contexts and as diverse as possible, adopting measures to improve data quality, and recording limitations of data via transparent documentation.[25] Informing humans across the lifecycle of biases in the data strengthens humans' ability to foresee limitations and impacts of the systems, as well as to "understand in which situation the performance of the model is the lowest". (1)

23 This set of practices involves Article 36 legal reviews for states that are party to the Additional Protocol I to the Geneva Conventions. The practices in this toolkit are not equivalent to states' legal obligations.

24 Rupert Barrett-Taylor and Natasha Karner, *AI Won't Replace the General: Algorithms, Decision-Making and Battlefield Command* (The Alan Turing Institute, 2025), https://www.turing.ac.uk/news/publications/ai-wont-replace-general-algorithms-decision-making-and-battlefield-command.

25 On biases in data see Alexander Blanchard and Laura Bruun, *Bias in Military Artificial Intelligence* (Stockholm International Peace Research Institute, 2024), https://doi.org/10.55163/CJFT9557; Holland Michel, *Known Unknowns: Data Issues and Military Autonomous Systems*.

H **Involve actors beyond technical specialists, including lawyers, ethicists, and operators.** Given that TEVV is not just a technical matter, the input of actors with ethical, legal, and operational expertise is needed to ensure that AI systems achieve the planned outcomes, while adhering to ethical and legal standards. This may involve blending the conduct of TEVV with the review of new weapons for compatibility with states' international legal compilations, as stipulated for states parties to the Additional Protocol I to the Geneva Conventions (Article 36 reviews).[26] It also includes continuously receiving feedback from operators. Such feedback can account for some unpredictable behaviours or situations and contribute to deciding on the necessary updates for supporting the exercise of human agency. (1, 2, 3)

Stage 5: Considering the human: Education, training, and human-system integration

Practices at this stage will mainly concern officials responsible for designing military training and actors involved in implementing measures to train about AI systems as well as human–AI interaction. While education and training encompass a specific stage of the lifecycle, they also designate sets of practices that are ongoing through the lifecycle in a non-linear way.

A **Train humans involved across the lifecycle on the objectives of AI systems set at earlier stages**. This practice ensures that there is alignment between the objectives of developing and using AI systems across the lifecycle via established lines of documentation between human actors. Training covers digital or tech literacy, which strengthens humans' understanding of the systems. A key element of training is learning about decision-making processes, organisational culture, political and institutional structures, as well as the roles and functions that AI systems play within those structures. This should be part of military personnel's comprehensive education to include strategic, operational, political, legal, ethical, and technical components and applies to all humans involved in the lifecycle. Informing humans about these aspects bolsters their ability to understand systems' functions within operational contexts and to adopt the necessary measures at each lifecycle stage. (1, 2)

B **Train operators to use the interface in a way that complies with objectives and limitations set previously**. This practice covers training about tech literacy and specifics of the interface. Elements of training include using the interface following the risk assessment frameworks established at earlier stages, employing potential mechanisms for intervention and backup measures, as well as conditions that might affect or degrade operators' ability to use the interface in compliance with objectives. For instance, an interface that does not constantly change or update assists with making human–AI interaction more predictable. Teaching standard operating procedures in conjunction with how the information is presented on the interface helps strengthening users' understanding of the system and how it may function in certain contexts and environments of use. (1)

C **Continuously train on cross-checking and critically assessing the outputs of AI systems.** This practice is an important part of ensuring that humans using the systems do not become complacent. It also allows human users to continuously sharpen their own assessments of AI systems' outputs, contributing to users' capacity to deliberate upon AI systems' outputs, as well as to safeguards around human deskilling. (2)

D **Continuously train on the dynamics of human–AI interaction.** Teaching humans across the lifecycle, including those in the operational environments, how to engage in critical reflections (including legal and ethical deliberations)

26 The practices in this toolkit are not equivalent to states' legal obligations.

during interactions with AI systems strengthens humans' awareness about dynamics of interactions with AI systems, such as automation bias. This practice entails providing hands-on experiences with the systems prior to use, including scenarios and simulations that would enhance humans' understanding of the affordances and limitations of using and interacting with AI systems, as well as their awareness of how to respond when systems do not perform as expected. Continuous training means updating education measures when needed, while considering educational and cultural backgrounds, and levels of digital literacy. Such training is particularly vital, but not limited to, senior military decision-makers. (1, 2)

E **Ensure a balance between sufficient understanding and appropriate calibration of trust in AI systems' outputs.** A balanced approach to training enables humans to understand systems' characteristics and functions, while remaining aware of limitations and adopting a critical perspective. This practice covers training on technical aspects and dynamics of human-AI interaction that may lead to either under-trust or over-trust of AI systems' outputs. Training measures aimed at relieving stress fatigue, decision fatigue, cognitive overload, or information overload, support operator resilience and users' ability to critically deliberate upon and assess AI outputs, especially when several AI systems are integrated in a network. Appropriate calibrations of trust boost users' ability to understand the system and deliberate on responses or courses of action. (1, 2)

F **Ensure ways for users to understand when systems are not working according to planned objectives and requirements**. This can be via built-in cues, safeguards and standards that strengthen the reliability of AI systems and humans' ability to respond in case of unexpected situations. This practice includes training humans about blind spots and edge cases to prevent over-trust and de-skilling. Training that instructs steps to follow in both foreseen/planned and unexpected situations strengthens humans' ability to understand and critically assess outputs of AI systems. This practice includes building awareness and strengthening knowledge of "machine-machine interaction", or of how machines "communicate with each other and coordinate their behaviour", especially as part of a complex network of AI systems. It also involves educating human personnel about detecting security threats, such as in case AI systems are hacked or jammed, as well as the safety risks of AI systems being compromised or subject to cyberattacks. Finally, the practice also extends to educating personnel about steps to follow in case of failures. (1, 2)

G **Adapt training measures dynamically when AI systems develop or are updated.** This practice recognises that training evolves in light of technological and operational changes. This applies to changes to the interfaces of the systems, changes in the wider ecosystem of technological elements that make up the AI systems, as well as modifications to the objectives of developing and using AI systems in relation to the intended context of use. (1)

Stage 6: Political and strategic considerations

Practices at this stage will mainly concern strategic and political decision-makers, as well as military officials. While political and strategic considerations encompass a specific stage of the lifecycle, they also designate sets of practices that are ongoing throughout the lifecycle in a non-linear way.

A **Raise awareness among political decision-makers about the benefits and limitations of AI systems without over-hyping or understating AI technologies**. This set of practices involves strengthening political and military officials' understanding of AI systems' characteristics and functioning within particular contexts, enabling them to make decisions about the objectives of developing and using AI systems. This includes officials' deliberation on when an "AI or an autonomous system is not the most appropriate tool" in a certain context or for a certain task. (1)

B **Incorporate concerns about the escalation risks of military applications of AI and misleading narratives amid geopolitical rivalry**. This practice explicitly invites decision-makers to consider how the development and use of AI systems relate to the proliferation of military technologies, escalation, or lowering the threshold for the use of force. It also covers considering the effects of reinforcing or elevating some narratives about AI and human agency, for example that AI inevitably increases efficiency and precision, and thereby enhances decision-makers' ability to deliberate upon the objectives and risks of developing and using AI systems. (1, 2)

C **Assess the conduct of practices within adversarial contexts**. The practices in this toolkit are not performed in isolation from broader societal and (geo-)political contexts. It is key for decision-makers to consider how the development and use of AI may be affected by the adversarial environment. This includes weighing potential adversarial behaviour and responses, including what they may imply in terms of reputational costs and relations with perceived allies and adversaries. This practice includes evaluating whether AI systems should be used in the applicable contexts, especially considering security risks such as unintended escalation (see stage 6, practice B). Assessing AI systems and related practices contributes to delimiting appropriate uses of AI systems within adversarial contexts and allowing users to deliberate upon courses of action to respond appropriately within those contexts. (2)

Stage 7: Operational-level command and control & Stage 8: Tactical employment

Practices at this stage will mainly concern users of AI systems, commanders at various levels, as well as operators.

The toolkit combines practices across stage 7 (operational-level command and control) and stage 8 (tactical employment) because there is a degree of overlap and repetition at those stages.

A **Ensure the deployment of AI systems is limited to anticipated circumstances and environments, to the extent possible.** Using AI systems within the intended environments increases the possibility of practices contributing to the exercise of human agency across the entire lifecycle aligning with the operational context. (1, 2, 3)

B **Follow safeguards and standards designed for users to understand if the systems do not work as planned**. This practice covers implementing the limitations and restrictions set at earlier stages, such as monitoring mechanisms and "real-time battlefield assessment to detect errors and tendencies toward error". Such safeguards allow human users to decide how to act in case of malfunctions, and respond in ways that affect the interaction with the system, for example, if an operator critically assesses the outputs of AI systems and pursues another course of action that is not suggested by these outputs. (2, 3)

C **Implement the risk assessment frameworks and mechanisms for human intervention based on the risk thresholds.** This practice ensures continuity of the risk assessment frameworks and limitations set at earlier stages. Mechanisms for human intervention include ways to slow down, manually override or deactivate the systems, reducing the tempo of use by adopting tactical patience, or having targeting decisions reviewed by multiple operators. Such mechanisms allow human users to critically assess AI systems' outputs and cross-check them with other sources of intelligence, strengthening their ability to deliberate upon actions and act in a way that may impact their use of and interaction with the systems. (2, 3)

D **Continuously evaluate the development of the systems during use and relaying feedback to TEVV.** This practice recognises that AI systems will likely be undergoing "tiny AI lifecycles" at the point of use through collecting new data and training the underlying machine learning models. Keeping track of these changes and updates involves additional evaluation and documentation. This evaluation strengthens human users' understanding of the systems and potential to deliberate upon actions, especially when combined with previous practices on risk assessment and mechanisms of human intervention. (1, 2)

E **Log the targeting process and create appropriate levels of documentation.** What constitutes appropriate levels of documentation depends on the use of AI systems in relation to the risk assessment framework, with systems evaluated as high-risk requiring more stringent documentation. Documenting the monitoring of AI systems during employment contributes to feedback mechanisms between users and developers. Such feedback loops, accompanied by organisational cultures prioritising honest exchange, in turn strengthen humans' understanding of the systems, as well as the systems' functions and limitations in certain contexts. (1)

F **Collect information during use, as in the case of 'black boxes' used in the civilian aviation context.** Such 'black boxes' collect data on malfunctions or accidents. They need not necessarily be physical devices but could rather take the form of network connections that update remotely accessible data storage. This practice would enable documenting events that happened during use, or information from the involved humans' viewpoints (not necessarily the technical data), including throughlines of accountability.[27] This practice requires forms of explainable or interpretable AI to document "the reasonings of the system both in the moment of decision-making and afterwards". This practice complements practice E of stages 7-8, as well as feedback loops between end-users and developers (see cross-cutting practice B). It contributes to strengthening humans' understanding of the systems and their functioning in certain contexts and environmens of use. (1)

27 See the ongoing activity 'responsibility, accountability, and knowledge transfers' in IEEE SA, *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*.

Stage 9: Review, reuse, and/or retire

Practices at this stage will mainly concern military commanders at different levels that have been previously involved in various contexts of deployment.

A **Implement reviews to evaluate whether previously established political, strategic, legal, ethical, and technical considerations have been met.** These comprehensive reviews contribute to updating objectives set at the earlier stages, risk assessment frameworks, and limitations applying to the development and use of AI systems. There are currently no transparency standards that address the design, data, and models used. This practice therefore promotes creating a systematic approach—a set of best practice guidelines—for reviewing and ensuring transparency. Reviews and updates are needed to ensure that humans across the lifecycle understand AI systems, have the potential to deliberate upon actions, and act in an impactful way.[28] (1, 2, 3)

B **Process data collected during use, compare the outcomes of the systems' use to risk assessment frameworks, and document whether the systems can be reused (in the case of hardware), actively used, paused, discontinued, or need to be updated.** This practice entails organising a multi-stakeholder panel involving military, legal, political, technical, and other experts allowing for "institutional learning and updating" the risk assessment frameworks applying to the development and use of AI systems. The practice may also cover using AI technologies to analyse the data collected during deployment and assist with tracing accountability for violations of legal and other frameworks. Documenting how the use of AI systems compared to risk assessment frameworks involves recording data pedigree and tracing the data lifecycle, or the lineage of the data, especially for potential reuse of the data. This documentation contributes to humans' understanding of the systems, as well as the systems' functions and limitations in certain contexts and environments of use. (1)

C **Communicate the findings to decision-makers and developers to support feedback loops.** This practice reiterates the importance of creating various opportunities for feedback loops between actors at different stages of the AI lifecycle. At this stage of the lifecycle, feedback loops that follow careful debriefing and lead to curated feedback are particularly important. In the context of AI systems, curated feedback implies reviewing feedback data that would be used for system updates. The practice may also cover employing AI technologies or automated systems to assist in analysing the data collected. Post-deployment feedback mechanisms, based on comprehensive reviews (stage 9, practice B), are key contributions to strengthening all three elements of human agency. (1, 2, 3)

D **Actively review and retest following system updates.** This practice underlines the need for continuous review processes once system training data or system architecture have changed. Such reviews strengthen humans' understanding of the functioning of AI systems. (1)

28 On iterative reviews and assessments see Jonathan Kwik, "Iterative Assessment for Military Artificial Intelligence (AI) Systems," in *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence*, ed. Berenice Boutin, Taylor Woodcock, and Sadjad Soltanzadeh (T.M.C. Asser Press, forthcoming 2026), https://ssrn.com/abstract=5338856.

Cross-cutting practices

These practices appear at every lifecycle stage in various forms; therefore, they apply to various human actors including designers, developers, testers, commanders, and users.

A Train on AI systems and human-AI interaction

Training is a continuous activity across the lifecycle applying to different human actors including political and strategic decision-makers, developers, engineers, testers, commanders, and other members of staff.

Training practices include various forms of hands-on experience with the systems, scenarios, wargames, or simulations which expose humans to the systems' functions and behaviours in certain contexts, as well as factors affecting the systems' performance not only in intended operational contexts but also "edge cases, ambiguous cases, and outright system failures" to reflect the unpredictability of the real world.

This set of cross-cutting practices encompasses not only training about the functioning of the systems, including their opportunities and limitations, but also training about the infrastructures and ecosystems of resources behind those systems, as well as dynamics of human-AI interaction and ways of addressing phenomena arising as part of these interactions, notably biases. Training about human-AI interaction raises humans' awareness about the potential impacts of using AI systems on human behaviour, human cognitive skills, and the exercise of human judgement.

Education programmes including wide-ranging implications of using AI systems allow humans, and especially users, to understand what legal and other frameworks are applicable and how to behave in alignment with these frameworks. Educating human actors about AI systems and the dynamics of human-AI interaction reinforces their ability to decide upon courses of actions in a timely manner. Training operators in the critical assessment of AI systems' outputs and the use of intervention mechanisms enables them to act in a way that impacts the use of the systems.

Note: Precise training measures will depend on the resources available to the organisations.

B Ensure communication and organisational learning

This set of practices underlines the importance of processes enabling communication among actors across the lifecycle in contributing to how these humans exercise their agency.

Feedback loops between end-users and actors at earlier stages enable decision-makers, developers, designers, testers, and other human actors to revise existing safeguards or implement new measures contributing to strengthening the exercise of human agency in intended contexts of use. Involving multiple actors including political decision-makers, developers, lawyers, ethicists, operators, and others allows these actors to build shared understandings of mechanisms to strengthen human agency across the lifecycle. In turn, this awareness strengthens humans' understanding of the particular AI systems and how they function in certain contexts, and building upon this, transmit knowledge that allows humans to act in a way that impacts the use of and interaction with the systems.

Different types of feedback mechanisms are used for different purposes. For example, some feedback loops can take the form of routine checks, others would focus on lessons learned from testing, while others would focus on feedback from use in a specific context—while not applying lessons learned from uses directly from one context to another.

As many defence organisations may acquire AI systems 'off the shelf' from the private sector, involving end-users in stages 1 and 2 may be challenging. Nevertheless, this involvement plays a key role in strengthening human agency across the lifecycle.

Note: There are various types of feedback and communication loops, and not all of them necessarily contribute to bolstering human agency. Moreover, this set of ongoing practices may be affected by the resources, infrastructure, and personnel available. The context (e.g., active armed conflict vs. peacetime) might also influence how this set of practices is implemented.

C Apply comprehensive risk assessment and mitigation

Designing a comprehensive risk assessment framework and following it throughout the lifecycle allows stakeholders to mitigate the multi-faceted risks (technical, legal, ethical, operational, and strategic) potentially arising from the development and use of AI systems in the military domain.[29]

A systematic overview of the risks that apply across the lifecycle provides human actors with the chance to strengthen their understanding of the AI systems and to deliberate upon how to address the risks. It also enhances their ability to act with an impact on the use of the systems in a way that is appropriate to the systems and the context. Therefore, a framework that is relevant to the systems and the resources available to the actors can contribute to the exercise of human agency.[30]

D Strengthen cybersecurity and reliability of AI systems

The exercise of human agency may be "impacted by the adversarial environment in which military decision-making occurs". This involves, for example, risks associated with adversarial attacks and various forms of hacking that are designed to manipulate the inputs of AI systems and thereby lead to corrupted, faulty, and destabilising outputs.[31] Other attacks might involve reverse-engineering the model, causing it to leak sensitive data from its underlying database to the adversary.

Designing and implementing cybersecurity measures aims at strengthening the reliability of AI systems, which in turn contributes to humans' ability to understand and reasonably foresee how the systems would function. Cybersecurity practices aim to ensure—to the extent possible—data security, robustness, and technical features that all contribute to humans' ability to respond adequately in the context of interacting with AI systems.[32] As one stakeholder highlights, rigorous TEVV practices throughout the lifecycle[33] strengthen the predictability and understandability of the systems and allow "greater human agency in the decisions that have been made on the basis of that system, regardless of where in the lifecycle they will be placed".

29 See the ongoing activity 'risk assessments' in IEEE SA, *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*.

30 On risk assessment frameworks see Assaad, "A Risk-Based Trust Framework for Assuring the Humans in Human-Machine Teaming"; Zena Assaad, "A Proposed Risk Categorisation Model for Human-Machine Teaming," *EICS '22: Engineering Interactive Computing Systems conference*, CEUR Workshop Proceedings, 2022, https://ceur-ws.org/Vol-3404/paper11.pdf; Blanchard et al., "A Risk-Based Regulatory Approach To Autonomous Weapon Systems."

31 Andrew Lohn, Anna Knack, Ant Burke, and Krystal Jackson, *Autonomous Cyber Defence: A Roadmap from Lab to Ops* (Center for Emerging Technology and Security & Center for Security and Emerging Security, 2023), https://cetas.turing.ac.uk/sites/default/files/2023-06/autonomous_cyber_defence_final_report.pdf.

32 For a legal perspective on analysing adversarial attacks see Jonathan Kwik, "Is Wearing These Sunglasses an Attack? Obligations under IHL Related to Anti-AI Countermeasures," *International Review of the Red Cross* 106, no. 926 (2024): 732–59, https://doi.org/10.1017/S1816383124000067.

33 See the ongoing activity 'TEVV, monitoring, hardware system or software updates and interoperability, maintenance' in IEEE SA, *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*.

Mechanisms that continuously monitor AI systems detect if systems function as intended and raise humans' awareness about potential malfunctions, thereby strengthening their understanding of systems' behaviours and their assessment of the context.

Note: This is a general set of practices at the intersection of AI technologies and cybersecurity measures that requires more spelling out. While this cannot be achieved in the context of this practical toolkit, we are flagging it as a point of further attention.

E Implement additional safeguards for online learning

The distinction between AI systems based on offline and online learning are significant in the context of sustaining and strengthening the exercise of human agency. Offline learning describes systems where learning happens only in the development phase, while for online learning systems, learning continues after deployment in operational contexts. [34]

Systems with online learning capabilities require additional safeguards contributing to the exercise of human agency because such systems will exhibit more unpredictable behaviour, with a higher risk for unintended consequences.

Additional safeguards regarding AI systems involving online learning are applied across the lifecycle. They may include closer monitoring of systems in operation, post-deployment reviews, as well as reviews prior to deployment, including Article 36 legal reviews.[35]

34 See Tim McFarland and Zena Assaad, "Legal Reviews of in Situ Learning in Autonomous Weapons," *Ethics and Information Technology* 25, no. 9 (2023), https://doi.org/10.1007/s10676-023-09688-9.

35 McFarland and Assaad, "Legal Reviews of in Situ Learning in Autonomous Weapons," 5. The practices in this toolkit are not equal to states' legal obligations.

Recommendations and take-aways

The toolkit of best practices identified by AutoPractices stakeholders as part of the co-creation process are already specific policy recommendations in and of themselves. In addition, we have summarised the various practices into the following four take-aways.

A Establish boundaries at the early stages of the AI lifecycle

Key practices to sustain and strengthen human agency need to be applied at early stages of the AI lifecycle.[36] Whatever happens at early lifecycle stages, including the human agency exercised there, affects what humans at later stages are able to do. This insight applies across all levels of deployment and extends to commanders and senior leaders who may think they have more agency than they actually do because of the decisions made earlier in the AI lifecycle.

Setting specific objectives and establishing boundaries for developing and using AI systems strengthen humans' understanding of how these systems work in particular contexts. It also provides mechanisms for humans to deliberate upon actions and creates opportunities to make an impact, thereby supporting all three elements of human agency: *1) understand and reasonably foresee a system's functions and effects in a relevant context; and 2) to deliberate and decide upon suitable actions in a timely manner; and 3) to act in a way that can impact the use of the system* (see Box 1).

Policies, doctrines and frameworks such as Use Cases and CONOPS should provide clear guidance at further stages of the lifecycle. Early stages of the lifecycle should focus on giving humans the opportunity to make as many "meaningful choices" as possible down the line. If some risks are left unaddressed, other practices, for example TEVV, might not be able to address these issues.

Boundaries and constraints set at early stages can be spatial, temporal, relating to the type of targets (e.g., prohibiting anti-personnel weapons), the targeting parameters, or to the operational environment.[37] They should reflect political, legal, ethical, societal considerations surrounding the integration of AI systems into the military domain. Such considerations should be informed by feedback from the end-users of AI systems, such as operators. The design of AI systems should be driven by the needs at the point of use, rather than the ambition of integrating AI for the sake of it or only based upon vendors' or software suppliers' claims about the capabilities of AI systems.

36 Ariel Conn and Ingvild Bode, "Establishing Human Responsibility and Accountability at Early Stages of the Lifecycle for AI-Based Defence Systems," *Ethics and Information Technology* 27, no. 51 (2025), https://doi.org/10.1007/s10676-025-09862-1.

37 On operational constraints see Nathan Wood, "Autonomous and AI-Enabled Systems: Extensions or Replacements of Human Will and Control?" *Ethics and Information Technology* 28, no. 3 (2026), https://doi.org/10.1007/s10676-025-09876-9.

At the procurement and acquisition stage, relevant actors need to conduct assessments related to the objectives set beforehand, especially when the procurement of commercially available systems is involved. Procurement officials need to factor in ways of addressing these contradictions.

B Implement feedback and communication mechanisms across the AI lifecycle

The design and development of AI systems in the military domain should be guided by communication and feedback loops between various human actors, especially users and developers. This is because the stages of the lifecycle are not mutually exclusive. The knowledge and awareness that humans have across the lifecycle about what happened at previous stages contribute to strengthening human agency and the overall oversight of AI systems.

Continuous communication and throughlines of documentation between actors contribute to transparency,[38] and thereby also the accountability of different human actors across the lifecycle.[39]

Documentation ensuring traceability includes practices such as pre-action constraints and post-deployment auditing, tracking human decisions across the lifecycle, communication and transparency, as well as awareness of practices performed at earlier stages. Yet, documentation itself does not mitigate risks: it needs to be considered in the context of different organisational cultures. Organisational cultures prioritising honest feedback are more likely to strengthen human agency. Without honest feedback, the risk of not sharing if something does not work increases, causing a cascading "domino effect" on the exercise of human agency. Organisational cultures prioritising action or speed may overlook the significance of key practices that contribute to exercising human agency.[40]

Post-deployment documentation contributes to ensuring legal accountability as well as revealing technical malfunctions and vulnerabilities.[41] Logging and recording data, as well as analysing documentation, to establish throughlines of accountability could be facilitated by dedicated AI tools. However, it is important to be aware of trade-offs between different approaches towards documentation, which each actor should navigate appropriately to the context.[42]

These feedback loops and communication mechanisms should be checked for both their relevance and reliability, making sure that feedback does not degrade the AI systems' performance.

38 On transparency see Branka Marijan, *Through a Glass, Darkly: Transparency and Military AI Systems*, CIGI Paper No. 315 (Centre for International Governance Innovation, 2025), https://www.cigionline.org/publications/through-a-glass-darkly-transparency-and-military-ai-systems/.

39 AutoPractices stakeholders highlight the need for humans to reasonably foresee a system's outcomes in a context in order to be held accountable for violations that might arise as a result of using that system.

40 Militaries would benefit from engaging in greater depth with the research findings and specific strategies in human-computer interaction happening across domains such as healthcare, transport, or mobility. In other words, sustaining and strengthening the exercise of human agency in the military domain would gain from developing insights that cut across other high-stakes, safety-critical domains of application.

41 See the ongoing activity 'responsibility, accountability, and knowledge transfers' in IEEE SA, *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*.

42 On trade-offs see Qiaochu Zhang, "Navigating the Complexities of Exercising Human Agency in Human-Machine Interaction Across the AI Lifecycle," *The AutoNorms Blog*, October 21, 2024, https://www.autonorms.eu/navigating-the-complexities-of-exercising-human-agency-in-human-machine-interaction-across-the-ai-lifecycle/.

C Prioritise education and training across the AI lifecycle

The education and training of human actors involved across the lifecycle should focus on providing sufficient knowledge of and confidence in AI systems, while raising awareness about AI systems' limitations, especially software-related issues and the dynamics of human-AI interaction. These training measures should be about examining the whole infrastructure and network involved. They should not just cover singular AI systems.[43]

While many issues related to AI and human-AI interaction may not be fully eliminated, training and awareness-raising contribute to reducing and mitigating their effects, as they allow humans to apply a critical lens to their interactions with AI systems. Training contributes to all elements of human agency, especially if it touches upon how humans interact with AI systems within a specific context and considers these humans' backgrounds.

Training of various humans along the lifecycle should also cover military ethics, and broader ethical and legal frameworks including IHL and international human rights law, paying specific attention to the impacts of human-AI interaction on the exercise of human judgement in relation to these frameworks.[44]

D Adopt strategies and steps to implement the practices listed in this toolkit

The stakeholders involved in the co-creation of this toolkit come from different backgrounds and therefore engage in various practices. This toolkit does not resolve these differences in experience and perspective but highlights the potential practices that could become solutions. The toolkit is intentionally diverse and diversified, underlining that there is no single, universal solution.

Therefore, many details surrounding practices in this toolkit will derive from the local contexts and conditions in which these systems are developed.[45] The implementation of the toolkit's practices will also depend on the resources available to organisations developing and using these technologies. It is up to stakeholders involved in the development and use of military systems integrating AI to adopt the necessary steps to elaborate specific practices according to context.

One first step in this process would be to take stock of the resources, infrastructure, and expertise available in house, as well as organisational learning processes, levels of bureaucracy, and the contexts informing the militaries (e.g., active conflict or peacetime). Additionally, this process could assess the military's engagement with the private sector and its reliance on infrastructure from industry, along with its priorities for interoperability with allies and the development of sovereign technologies and systems, to the extent possible.

Based on the first step, a second step would be to organise reviews of current practices, their alignment with this toolkit, and selecting priority practices

43 See the the ongoing activity 'considering the human: training, education, and human-system integration' in IEEE SA, *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*.

44 For a discussion on dynamics of human-machine interaction and what they mean for compliance with IHL principles, see Klaudia Klonowska and Taylor Woodcock, "Rhetoric and Regulation: The (Limits of) Human/AI Comparison in Legal Debates on Military AI" in *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence*, ed. Berenice Boutin, Taylor Woodcock, and Sadjad Soltanzadeh (T.M.C. Asser Press, forthcoming 2026), https://ssrn.com/abstract=5417097.

45 See Alexander Blanchard, Christopher Thomas, and Mariarosaria Taddeo, "Ethical Governance of Artificial Intelligence for Defence: Normative Tradeoffs for Principle to Practice Guidance," *AI & SOCIETY* 40 (2025): 185–98, https://doi.org/10.1007/s00146-024-01866-7.

(in addition to those that are legal obligations, which are a must), as well as practices to pursue, update, or change. A third step could involve an exchange of practices within a multistakeholder format focused on capacity building.[46] While practices related to capacity building are outside the scope of this toolkit, we recommend that stakeholders design concrete strategies not only to implement the toolkit, but also to address potential barriers to implementation. This may include developing strategies to engage with the private sector and suppliers of both hardware and software in the context of military applications of AI.

We further recommend that policymakers, developers, and users of military AI systems articulate concrete steps to implement practices in accordance with the intended contexts of use and the resources available to these actors. Taking such steps demonstrates a commitment to normalising a positive version of human agency via bottom-up approaches to the governance of AI in the military domain.

46 See core recommendations 4 and 5 in The Global Commission on Responsible Artificial Intelligence in the Military Domain, *Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the MIlitary Domain* (The Hague Centre for Strategic Studies, 2025), https://hcss.nl/wp-content/uploads/2025/09/GC-REAIM-Strategic-Guidance-Report-Final-WEB.pdf.

Conclusion

This practical toolkit is the capstone output of the ERC-funded AutoPractices project, which demonstrates the need for a bottom-up approach towards governing applications of AI in the military domain.

It encompasses a total of 54 sets of best practices contributing to sustaining and strengthening the exercise of human agency, as identified by 49 stakeholders representing various fields of expertise and regions of the world. These practices can be performed by, and are applicable to, many actors across all stages of the lifecycle of AI systems.

The toolkit of practices all along the lifecycle is the key recommendation put forward by the AutoPractices project. It addresses various actors along the AI lifecycle in the military domain. These practices can be summed up into the following four key take-aways to sustain and strengthen human agency in relation to military applications of AI:

1. **Establish boundaries at the early stages of the AI lifecycle**
2. **Implement feedback and communication mechanisms across the AI lifecycle**
3. **Prioritise education and training across the AI lifecycle**
4. **Adopt strategies and steps to implement the practices listed in this toolkit**

With this practical toolkit, we call upon other stakeholders in the debate on the integration of AI into military systems to further pursue bottom-up ways of practically implementing high-level and top-down principles. At a time when militaries are increasingly investing into the development of AI technologies while also putting forward strategies towards the 'responsible' development and use of AI in defence, understanding how the exercise of human agency can be strengthened in practice is particularly crucial.

References

List of publications by the AutoPractices team (selected)[47]

Blanchard, Alexander, and Laura Bruun. 2024. *Bias in Military Artificial Intelligence*. Stockholm International Peace Research Institute. https://doi.org/10.55163/CJFT9557.

Blanchard, Alexander, and Laura Bruun. 2025. *Autonomous Weapon Systems and AI-Enabled Decision Support Systems in Military Targeting: A Comparison and Recommended Policy Responses*. Stockholm International Peace Research Institute. https://doi.org/10.55163/YQBY3151.

Blanchard, Alexander, and Netta Goussac. 2025. *Towards Multilateral Policy on Autonomous Weapon Systems*. Stockholm International Peace Research Institute. https://doi.org/10.55163/VLLZ7787.

Blanchard, Alexander, Claudio Novelli, Luciano Floridi, and Mariarosaria Taddeo. 2025. "A Risk-Based Regulatory Approach To Autonomous Weapon Systems." *Digital Society* 4 (23). https://doi.org/10.1007/s44206-025-00181-y.

Blanchard, Alexander, and Mariarosaria Taddeo. 2023. "The Ethics of Artificial Intelligence for Intelligence Analysis: A Review of the Key Challenges with Recommendations." *Digital Society* 2 (12). https://doi.org/10.1007/s44206-023-00036-4.

Blanchard, Alexander, Christopher Thomas, and Mariarosaria Taddeo. 2025. "Ethical Governance of Artificial Intelligence for Defence: Normative Tradeoffs for Principle to Practice Guidance." *AI & SOCIETY* 40: 185–98. https://doi.org/10.1007/s00146-024-01866-7.

Bode, Ingvild. 2023. "Practice-Based and Public-Deliberative Normativity: Retaining Human Control over the Use of Force." *European Journal of International Relations* 29 (4): 990–1016. https://doi.org/10.1177/13540661231163392.

Bode, Ingvild. 2025. *Emerging Norms around Military Applications of AI: The Case of Human Control*. GC REAIM Expert Policy Note Series. The Hague Centre for Strategic Studies. https://hcss.nl/wp-content/uploads/2025/05/Bode-1.pdf.

Bode, Ingvild. 2025. *Human-Machine Interaction and Human Agency in the Military Domain*. Policy Brief No. 193. Centre for International Governance Innovation. https://www.cigionline.org/publications/human-machine-interaction-and-human-agency-in-the-military-domain/.

Bode, Ingvild, **Hendrik Huelss**, **Anna Nadibaidze**, **Guangyu Qiao-Franco**, and **Tom F.A. Watts**. 2023. "Prospects for the Global Governance of Autonomous Weapons: Comparing Chinese, Russian, and US Practices." *Ethics and Information Technology* 25 (5). https://doi.org/10.1007/s10676-023-09678-x.

47 The names of the AutoPractices team members are in bold.

Bode, Ingvild, **Anna Nadibaidze**, **Tom F.A. Watts**, and **Qiaochu Zhang**. 2025. "Ensuring the Exercise of Human Agency in AI-Based Military Systems: Concerns across the Lifecycle." *Ethics and Information Technology* 27 (50). https://doi.org/10.1007/s10676-025-09861-2.

Bode, Ingvild, and **Tom F.A. Watts**. 2021. *Meaning-Less Human Control: Lessons from Air Defence Systems on Meaningful Human Control for the Debate on AWS*. Drone Wars UK & Center for War Studies. https://dronewars.net/2021/02/19/meaning-less-human-control-lessons-from-air-defence-systems-for-lethal-autonomous-weapons/.

Bode, Ingvild, and **Tom F.A. Watts**. 2023. *Loitering Munitions and Unpredictability: Autonomy in Weapon Systems and Challenges to Human Control*. Center for War Studies & Royal Holloway Centre for International Security. https://www.autonorms.eu/loitering-munitions-and-unpredictability-autonomy-in-weapon-systems-and-challenges-to-human-control/.

Conn, Ariel. 2025. "The Role of 'Everyone Else' in the International Effort Toward Responsible AI." *Perry World House*, November 24. https://perryworldhouse.upenn.edu/news-and-insight/the-role-of-everyone-else-in-the-international-effort-toward-responsible-ai/.

Conn, Ariel, Zena Assaad, and Catherine Tessier. 2025. *An Approach for Assessing Autonomous and AI-Enabled Capabilities within Weapons Systems*. GC REAIM Expert Policy Note Series. The Hague Centre for Strategic Studies. https://hcss.nl/wp-content/uploads/2025/05/Conn-et-al.pdf.

Conn, Ariel, and **Ingvild Bode**. 2025. "Establishing Human Responsibility and Accountability at Early Stages of the Lifecycle for AI-Based Defence Systems." *Ethics and Information Technology* 27 (51). https://doi.org/10.1007/s10676-025-09862-1.

Nadibaidze, Anna. 2024. "Do AI Decision Support Systems 'Support' Humans in Military Decision-Making on the Use of Force?" *Opinio Juris*, November 29. https://opiniojuris.org/2024/11/29/do-ai-decision-support-systems-support-humans-in-military-decision-making-on-the-use-of-force/.

Nadibaidze, Anna, **Ingvild Bode**, and **Alexander Blanchard**. 2025. "From Human Control to Human Agency in the Debate about Military Applications of Artificial Intelligence." Preprint, *SSRN*. https://doi.org/10.2139/ssrn.5755743.

Nadibaidze, Anna, **Ingvild Bode**, and **Qiaochu Zhang**. 2024. *AI in Military Decision Support Systems: A Review of Developments and Debates*. Center for War Studies. https://www.autonorms.eu/ai-in-military-decision-support-systems-a-review-of-developments-and-debates/.

The AutoNorms Project. 2023. "Global Governance of AI in the Military Domain." September 28. https://www.autonorms.eu/global-governance-of-ai-in-the-military-domain/.

The AutoPractices Project. 2025. *Map of Practices*. Center for War Studies. https://www.autonorms.eu/governing-ai-technologies-in-the-military-domain-from-the-bottom-up-map-of-practices/.

Zhang, Qiaochu. 2024. "Navigating the Complexities of Exercising Human Agency in Human-Machine Interaction Across the AI Lifecycle." *The AutoNorms Blog*, October 21. https://www.autonorms.eu/navigating-the-complexities-of-exercising-human-agency-in-human-machine-interaction-across-the-ai-lifecycle/.

List of publications by the AutoPractices stakeholders[48]

Assaad, Zena. 2022. "A Proposed Risk Categorisation Model for Human-Machine Teaming." *EICS '22: Engineering Interactive Computing Systems conference*, CEUR Workshop Proceedings, June 21, 1–6. https://ceur-ws.org/Vol-3404/paper11.pdf.

Assaad, Zena. 2024. "A Risk-Based Trust Framework for Assuring the Humans in Human-Machine Teaming." *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, September 16, 1–9. https://doi.org/10.1145/3686038.3686045.

Assaad, Zena, and **Adam Hepworth**. 2025. *A Systems Engineering Lifecycle Approach to Responsible AI*. GC REAIM Expert Policy Note Series. The Hague Centre for Strategic Studies. https://hcss.nl/wp-content/uploads/2025/05/Zena-Assaad-and-Adam-Hepworth.pdf.

Barrett-Taylor, Rupert, and Natasha Karner. 2025. *AI Won't Replace the General: Algorithms, Decision-Making and Battlefield Command*. The Alan Turing Institute. https://www.turing.ac.uk/news/publications/ai-wont-replace-general-algorithms-decision-making-and-battlefield-command.

Bhila, Ishmael. 2024. "Putting Algorithmic Bias on Top of the Agenda in the Discussions on Autonomous Weapons Systems." *Digital War* 5: 201–12. https://doi.org/10.1057/s42984-024-00094-z.

Bode, Ingvild, and **Ishmael Bhila**. 2024. "The Problem of Algorithmic Bias in AI-Based Military Decision Support Systems." *ICRC Humanitarian Law & Policy Blog*, September 3. https://blogs.icrc.org/law-and-policy/2024/09/03/the-problem-of-algorithmic-bias-in-ai-based-military-decision-support-systems/.

Bryson, Joanna J. 2020. "The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.1.

Cawthorne, Dylan. 2023. *The Ethics of Drone Design: How Value-Sensitive Design Can Create Better Technologies*. Routledge.

Cawthorne, Dylan, and Arne Devos. 2020. "Capability Caution in UAV Design." *2020 International Conference on Unmanned Aircraft Systems (ICUAS)*, IEEE, 1572–81. https://doi.org/10.1109/ICUAS48674.2020.9214008.

Cawthorne, Dylan, and Aimee Robbins-van Wynsberghe. 2020. "An Ethical Framework for the Design, Development, Implementation, and Assessment of Drones Used in Public Healthcare." *Science and Engineering Ethics* 26 (5): 2867–91. https://doi.org/10.1007/s11948-020-00233-1.

Cole, August, Dan Howard, Robert Latiff, George Lucas, **Gregory M. Reichberg**, and Kaushik Roy. 2024. *Artificial Intelligence in Military Planning and Operations*. Peace Research Institute Oslo. https://www.prio.org/publications/13815.

Dorsey, Jessica. forthcoming 2026. "The Erosion of Human(e) Judgement in Targeting? Quantification Logics, AI-Enabled Decision-Support Systems and Proportionality Assessments in IHL." *International Review of the Red Cross*. https://ssrn.com/abstract=5910851.

48 Only publications by stakeholders who gave explicit consent to be named are included in this list. The names of stakeholders are in bold.

Dorsey, Jessica, and **Zena Assaad**. 2025. "Designing Lawful Military AI: Technical and Legal Reflections on Decision-Support and Autonomous Weapon Systems." *Perry World House*, November 24. https://perryworldhouse.upenn.edu/news-and-insight/designing-lawful-military-ai-technical-and-legal-reflections-on-decision-support-and-autonomous-weapon-systems/.

Dorsey, Jessica, and Marta Bo. forthcoming 2026. "AI-Enabled Decision-Support Systems in the Joint Targeting Cycle: Legal Challenges, Risks, and the Human(e) Dimension." *International Law Studies* 107. https://digital-commons.usnwc.edu/ils/vol107/iss1/7/.

Erskine, Toni. 2024. "AI and the Future of IR: Disentangling Flesh-and-Blood, Institutional, and Synthetic Moral Agency in World Politics." *Review of International Studies* 50 (3): 534–59. https://doi.org/10.1017/S0260210524000202.

Erskine, Toni. 2024. "Before Algorithmic Armageddon: Anticipating Immediate Risks to Restraint When AI Infiltrates Decisions to Wage War." *Australian Journal of International Affairs* 78 (2): 175–90. https://doi.org/10.1080/10357718.2024.2345636.

Erskine, Toni, and Steven E. Miller. 2024. "AI and the Decision to Go to War: Future Risks and Opportunities." *Australian Journal of International Affairs* 78 (2): 135–47. https://doi.org/10.1080/10357718.2024.2349598.

Evans, Katie D., Scott A. Robbins, and **Joanna J. Bryson**. 2025. "Do We Collaborate with What We Design?" *Topics in Cognitive Science* 17 (2): 392–411. https://doi.org/10.1111/tops.12682.

Ferey, Amélie, and **Laure de Roucy-Rochegonde**. 2024. "De l'Ukraine à Gaza : l'intelligence artificielle en guerre." *Politique étrangère* N° 243 (3): 39–50. https://doi.org/10.3917/pe.243.0039.

Garcia, Denise. 2023. *The AI Military Race: Common Good Governance in the Age of Artificial Intelligence*. Oxford University Press.

Garcia, Denise. 2025. "The Global Diplomacy of Governing Military Artificial Intelligence." *Ethics and Information Technology* 27 (55). https://doi.org/10.1007/s10676-025-09863-0.

Goussac, Netta. 2025. "Responsible Behaviour in Military AI Starts with Responsible Procurement." *Stockholm International Peace Research Institute*, October 16. https://www.sipri.org/commentary/essay/2025/military-ai-responsible-procurement.

Goussac, Netta, and Rain Liivoja. 2025. "CyCon 2025 Series – Legal Reviews of Military Artificial Intelligence Capabilities." *Lieber Institute West Point Articles of War*, August 25. https://lieber.westpoint.edu/legal-reviews-military-artificial-intelligence-capabilities/.

Hoffberger-Pippan, Elisabeth. 2021. "Autonomous Weapon Systems and Human Control: Politically Desired or also Legally Required?" In *Max Planck Yearbook of United Nations Law,* edited by Erika de Wet and Katharin Maria Scherr. Brill. https://doi.org/10.1163/18757413_02401009.

Hoffberger-Pippan, Elisabeth. forthcoming 2026. "ChatGPT for the Military? Large Language Models in the Military Domain the Role of Article 36 Weapons Reviews." In *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence*, edited by Berenice Boutin, Taylor Kate Woodcock, and Sadjad Soltanzadeh. T.M.C Asser Press.

Hoffberger-Pippan, Elisabeth. 2025. "Hypersonic Weapons and International Law: Incompatibilities with Arms Control Treaties and International Humanitarian Law." In *International Law and Technological Change: Testing the Adaptability of International Law*, edited by Irene Couzigou. Edward Elgar.

Holland Michel, Arthur. 2021. *Known Unknowns: Data Issues and Military Autonomous Systems*. United Nations Institute for Disarmament Research. https://unidir.org/known-unknowns.

Holland Michel, Arthur. 2024. *Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making*. International Committee of the Red Cross. https://shop.icrc.org/decisions-decisions-decisions-computation-and-artificial-intelligence-in-military-decision-making-pdf-en.html.

Holland Michel, Arthur. forthcoming 2026. *Autonomous Refusal in Lethal Weapons*. Center for War Studies, University of Southern Denmark.

Kaufhold, Marc-André, **Thea Riebe**, Markus Bayer, and Christian Reuter. 2024. "'We Do Not Have the Capacity to Monitor All Media': A Design Case Study on Cyber Situational Awareness in Computer Emergency Response Teams." *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, May 11, 1–16. https://doi.org/10.1145/3613904.3642368.

Kgomosotho, Keketso. 2025. *New and Emerging Terminologies in Ethical AI Principles: Exploring the International Law Implications in the Military Context*. GC REAIM Expert Policy Note Series. The Hague Centre for Strategic Studies. https://hcss.nl/wp-content/uploads/2025/05/Kgomosotho.pdf.

Kirkham, Joanne. 2022. *Les systèmes d'armes létaux autonomes : une étude de droit international*. Université Paris Panthéon-Assas. https://theses.fr/2022ASSA0068.

Klonowska, Klaudia. 2022. "Article 36: Review of AI Decision-Support Systems and Other Emerging Technologies of Warfare." In *Yearbook of International Humanitarian Law, Volume 23*, edited by Terry D. Gill, Robin Geiß, Heike Krieger, and Rebecca Mignot-Mahdavi. T.M.C. Asser Press. https://doi.org/10.1007/978-94-6265-491-4_6.

Klonowska, Klaudia, and Taylor Kate Woodcock. forthcoming 2025. "Rhetoric and Regulation: The (Limits of) Human/AI Comparison in Legal Debates on Military AI." In *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence*, edited by Berenice Boutin, Taylor Kate Woodcock, and Sadjad Soltanzadeh. T.M.C. Asser Press. https://ssrn.com/abstract=5417097.

Klonowska, Klaudia, and Sofie van der Maarel. 2025. "Deep Sensing, the New Military AI? Why Deep Sensing Should Be on the Radar." *Opinio Juris*, May 6. https://opiniojuris.org/2025/05/06/deep-sensing-the-new-military-ai-why-deep-sensing-should-be-on-the-radar/.

Knack, Anna, Bowden Alastair, Natasha Karner, et al. 2025. *Defence AI Assurance: Identifying Promising Practice and a System Card Template for Defence*. The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2025-09/defence_ai_assurance.pdf.

Knack, Anna, Nandita Balakrishnan, and Timothy Clancy. 2025. *Applying AI to Strategic Warning*. Centre for Emerging Technology and Security. https://cetas.turing.ac.uk/publications/applying-ai-strategic-warning.

Kozlovski, Atay. 2024. "When Algorithms Decide Who Is a Target: IDF's Use of AI in Gaza." *Tech Policy Press*, May 13. https://www.techpolicy.press/when-algorithms-decide-who-is-a-target-idfs-use-of-ai-in-gaza/.

Kozlovski, Atay. 2025. "Reasons Underdetermination in Meaningful Human Control." *Ethics and Information Technology* 27 (59). https://doi.org/10.1007/s10676-025-09858-x.

Kozlovski, Atay. forthcoming 2026. "Meaningful Human Control over AI Military Decision Support Systems: Exploring Key Challenges." In *Artificial Intelligence and Big Data Ethics in Military and Humanitarian Healthcare*, edited by Bernhard Koch and David Winkler. Military and Humanitarian Health Ethics. Springer Cham.

Kwik, Jonathan. 2022. "A Practicable Operationalisation of Meaningful Human Control." *Laws* 11 (3). https://doi.org/10.3390/laws11030043.

Kwik, Jonathan. 2024. *Lawfully Using Autonomous Weapon Technologies*. T.M.C. Asser Press.

Kwik, Jonathan. forthcoming 2026. "Iterative Assessment for Military Artificial Intelligence (AI) Systems." In *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence*, edited by Berenice Boutin, Taylor Kate Woodcock, and Sadjad Soltanzadeh. T.M.C. Asser Press. https://ssrn.com/abstract=5338856.

Lewis, Dustin A. 2022. "On 'Responsible AI' in War: Exploring Preconditions for Respecting International Law in Armed Conflict." In *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*, edited by Oliver Mueller, Philipp Kellmeyer, Silja Voeneky, and Wolfram Burgard. Cambridge University Press. https://doi.org/10.1017/9781009207898.037.

Lewis, Dustin A. 2023. "War Crimes Involving Autonomous Weapons: Responsibility, Liability and Accountability." *Journal of International Criminal Justice* 21 (5): 965–80. https://doi.org/10.1093/jicj/mqad027.

Lewis, Dustin A., and Hannah Sweeney. 2025. *Exercising Cognitive Agency: A Legal Framework Concerning Natural and Artificial Intelligence in Armed Conflict*. Harvard Law School Program on International Law and Armed Conflict. https://pilac.law.harvard.edu/exercising-cognitive-agency.

Lohn, Andrew, **Anna Knack**, Ant Burke, and Krystal Jackson. 2023. *Autonomous Cyber Defense: A Roadmap from Lab to Ops*. Centre for Emerging Technology and Security & Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/autonomous-cyber-defense/.

Marijan, Branka. 2025. *Through a Glass, Darkly: Transparency and Military AI Systems*. CIGI Paper No. 315. Centre for International Governance Innovation. https://www.cigionline.org/publications/through-a-glass-darkly-transparency-and-military-ai-systems/.

Marijan, Branka. 2025. *Trust by Design? AI in Military Applications*. Policy Brief No. 2018. Centre for International Governance Innovation. https://www.cigionline.org/publications/trust-by-design-ai-in-military-applications/

Marijan, Branka. forthcoming 2026. "AI in Warfare: Navigating the Ethical Minefield." In *Foundational Principles of AI Governance and Policy*, edited by Ann Fitz-Gerald, Vijay Ganesh, Jatin Nathwani, Maral Niazi, and Münür Sacit Herdem. Springer.

McFarland, Tim, and **Zena Assaad**. 2023. "Legal Reviews of in Situ Learning in Autonomous Weapons." *Ethics and Information Technology* 25 (9). https://doi.org/10.1007/s10676-023-09688-9.

Puscas, Ioana. 2022. *Human-Machine Interfaces in Autonomous Weapon Systems*. United Nations Institute for Disarmament Research. https://unidir.org/publication/human-machine-interfaces-in-autonomous-weapon-systems/.

Puscas, Ioana. 2023. *AI and International Security: Understanding the Risks and Paving the Path for Confidence-Building Measures*. United Nations Institute for Disarmament Research. https://unidir.org/publication/ai-and-international-security-understanding-the-risks-and-paving-the-path-for-confidence-building-measures/.

Puscas, Ioana. 2024. *Large Language Models and International Security: A Primer*. United Nations Institute for Disarmament Research. https://unidir.org/publication/large-language-models-and-international-security-a-primer/.

Qi, Haotian, and Wanmingkhwan Kamnerdrat. 2025. "How Epistemic Community Shapes Global Governance of AI in Military Domain." *Chinese Journal of International Politics* 18 (1): 85–122.

Reichberg, Gregory M., and Henrik Syse. 2024. "Ethical Analysis of AI-Based Systems for Military Use." In *Artificial Intelligence, Ethics and the Future of Warfare: Global Perspectives*, edited by Kaushik Roy. Routledge.

Riebe, Thea, Anja-Liisa Gonsior, Lilian Reichert, and Christian Reuter. 2025. "Envisioning Human-Machine Interaction in Future Warfare: Defence Industry Narratives on Human Control of Autonomous Weapon Systems." *Global Society* 39 (4): 421–45. https://doi.org/10.1080/13600826.2024.2436966.

Riebe, Thea, Stefka Schmid, and Christian Reuter. 2024. "Dual-Use Information Technology: Research, Development and Governance." In *Information Technology for Peace and Security: IT Applications and Infrastructures in Conflicts, Crises, War, and Peace*, 2nd ed., edited by Christian Reuter. Springer.

de Roucy-Rochegonde, Laure. 2024. *La guerre à l'ère de l'intelligence artificielle : quand les machines prennent les armes*. PUF.

de Roucy-Rochegonde, Laure. 2025. *Artificial Promises or Real Regulation? Inventing Global AI Governance.* French Institute for International Affairs (IFRI). http://www.ifri.org/en/studies/artificial-promises-or-real-regulation-inventing-global-ai-governance.

Schmitz, Chris, and **Joanna Bryson**. 2025. "A Moral Agency Framework for Legitimate Integration of AI in Bureaucracies (Extended Abstract)." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8 (3): 2292–93. https://doi.org/10.1609/aies.v8i3.36714.

Shanahan, Jack. 2024. "Symposium on Military AI and the Law of Armed Conflict: A Risk Framework for AI-Enabled Military Systems." *Opinio Juris*, April 1. https://opiniojuris.org/2024/04/01/symposium-on-military-ai-and-the-law-of-armed-conflict-a-risk-framework-for-ai-enabled-military-systems/.

Shanahan, Jack. 2025. "Artificial Intelligence and Nuclear Command and Control: It's Even More Complicated Than You Think." *Arms Control Association*, September. https://www.armscontrol.org/act/2025-09/features/artificial-intelligence-and-nuclear-command-and-control-its-even-more.

Shanahan, Jack. 2025. *The Need for A Lifecycle Approach to AI-Enhanced Autonomy*. Defense Paper Series. Special Competitive Studies Project. https://www.scsp.ai/wp-content/uploads/2025/09/Shanahan_Defense-Paper-Series.pdf.

Viveros Álvarez, Jimena Sofía. 2024. "Symposium on Military AI and the Law of Armed Conflict: Drone Swarms as Weapons of Mass Destruction." *Opinio Juris*, April 5. https://opiniojuris.org/2024/04/05/symposium-on-military-ai-and-the-law-of-armed-conflict-drone-swarms-as-weapons-of-mass-destruction/.

Viveros Álvarez, Jimena Sofía. 2024. "The Risks and Inefficacies of AI Systems in Military Targeting Support." *ICRC Humanitarian Law & Policy Blog*, September 4. https://blogs.icrc.org/law-and-policy/2024/09/04/the-risks-and-inefficacies-of-ai-systems-in-military-targeting-support/.

Viveros Álvarez, Jimena Sofía. 2024. "Why Should the UN 'Govern AI for Humanity': What Is at Stake and What Is the Urgency?" *Opinio Juris*, November 27. https://opiniojuris.org/2024/11/27/why-should-the-un-govern-ai-for-humanity-what-is-at-stake-and-what-is-the-urgency/.

Wood, Nathan Gabriel. 2024. "Explainable AI in the Military Domain." *Ethics and Information Technology* 26 (29). https://doi.org/10.1007/s10676-024-09762-w.

Wood, Nathan Gabriel. 2026. "Autonomous and AI-Enabled Systems: Extensions or Replacements of Human Will and Control?" *Ethics and Information Technology* 28 (3). https://doi.org/10.1007/s10676-025-09876-9.

Wood, Nathan Gabriel. forthcoming. "AI, Trust, and Teaming: The Humans-as-Handlers Approach for Autonomous and Opaque AI." *Minds and Machines*.

Appendices

Each of the four appendices serves a distinct purpose. Appendix A underlines why sustaining and strengthening the exercise of human agency in military decision-making with AI systems is important. This section offers in-depth thinking around the AutoPractices project's motivation by outlining six reasons for why the exercise of human agency along the whole entire AI lifecycle matters. Appendix B explains how the AutoPractices project, as a bottom-up effort to govern military applications of AI—and especially AI technologies in the use of forcre—contributes to the broader international discussions taking place in this field, which are mostly top-down. Appendix C describes the methodology behind AutoPractices' co-creative process and delineates the various project phases. Appendix D lists the questions used in stakeholder surveys and interviews, along with the scenario created by the AutoPractices team for use during the stakeholder workshops.

A The importance of exercising human agency in the use of force

The role of human involvement in targeting decision-making is a prominent topic in academic and policy discussions about military applications of AI.[49] While recognising the complexity of these ongoing debates, the AutoPractices project takes as its starting point the assumption that the exercise of human agency in the development and deployment of AI systems should be sustained and strengthened across these systems' lifecycle.[50]

One characteristic differentiating AI systems from other military systems is the importance of ensuring the exercise of human agency at the early stages of the lifecycle, as practices at the beginning matter for humans exercising agency at later stages. As one AutoPractices collaborator put it, if human agency is not exercised at earlier stages, "at the use stage, it might be too late". Exercising a high quality of human agency—the agency of various groups of humans involved across different lifecycle stages—is important for at least six reasons.

First, legal frameworks that apply to armed conflict, principally international humanitarian law (IHL), are centred around human actors as "only natural persons are capable of administering the performance of IHL obligations binding on States".[51] As AutoPractices stakeholders highlight, exercising human agency is necessary to ensure legal accountability of human actors for breaches of international law. Exercising human agency is thus "the basis for legal compliance", as well as for compliance more broadly.[52]

49 Ingvild Bode, *Emerging Norms around Military Applications of AI: The Case of Human Control*, GC REAIM Expert Policy Note Series (The Hague Centre for Strategic Studies, 2025), https://hcss.nl/wp-content/uploads/2025/05/Bode-1.pdf.

50 Ingvild Bode, Anna Nadibaidze, Tom F.A. Watts, and Qiaochu Zhang, "Ensuring the Exercise of Human Agency in AI-Based Military Systems: Concerns across the Lifecycle," *Ethics and Information Technology* 27, no. 50 (2025), https://doi.org/10.1007/s10676-025-09861-2.

51 Dustin A. Lewis and Hannah Sweeney, *Exercising Cognitive Agency: A Legal Framework Concerning Natural and Artificial Intelligence in Armed Conflict* (Harvard Law School Program on International Law and Armed Conflict, 2025), 10, https://pilac.law.harvard.edu/exercising-cognitive-agency.

52 See Arthur Holland Michel, *Autonomous Refusal in Lethal Weapons* (Center for War Studies, University of Southern Denmark, forthcoming 2026).

Second, exercising agency allows human actors to engage in the ethical and moral reasoning and judgments that are needed to prevent the risk of the technological systemisation of violence associated with AI technologies in warfare.[53] As warfare is a human activity that involves moral dilemmas, humans≈are expected to engage in ethical decision-making, especially in relation to decisions with severe potential consequences such as death, destruction, and human rights violations.

Third, the exercise of human agency contributes to mitigating issues arising as part of human-AI interaction. In particular, there is a need to strengthen human agency by being aware of and mitigating various human biases such as automation, cognitive, or action biases. This should also include being attentive to sycophancy on the part of the AI systems, or trust-related dynamics (whether undertrust or overtrust).[54] If left unaddressed, such challenges can raise significant security and humanitarian risks for both civilians and combatants. [55]

Fourth, sustaining and strengthening human agency is also strategically beneficial to military actors. Bolstering human agency at every level of the chain of command contributes to ensuring that military operations achieve their broader strategic goals, while complying with the necessary legal and ethical requirements in pursuit of these goals. As defence spending is increasing around the world, many actors pursue technological innovation, including in the sphere of AI technologies. Yet, innovation in military applications of AI need not be in opposition to bolstering the role of human. On the contrary, they can complement each other in a way that is compliant with strategic objectives, the law of armed conflict and other relevant norms, as well as appropriate ethical decision-making about the consequences of operations.

Fifth, in light of unsolved problems in AI research, sustaining and strengthening the exercise of human agency in high-stakes decision-making contexts such as the military domain is an operational and security necessity. There are questions about the reliability and robustness of AI systems. Certain types of AI applications, such as large language models (LLMs), entail limits to explainability and predictability. For example, LLMs inevitably produce factually incorrect information as part of their statistical core design.[56] Such AI applications therefore involve unexpected and unpredictable outputs, with security implications.[57] Concerns around reliability and predictability mean that

53 Neil C. Renic and Elke Schwarz, "Crimes of Dispassion: Autonomous Weapons and the Moral Challenge of Systematic Killing," *Ethics & International Affairs* 37, no. 3 (2023): 333, https://doi.org/10.1017/S0892679423000291; Toni Erskine, "AI and the Future of IR: Disentangling Flesh-and-Blood, Institutional, and Synthetic Moral Agency in World Politics," *Review of International Studies* 50, no. 3 (2024): 534–59, https://doi.org/10.1017/S0260210524000202.

54 Ishmael Bhila, "Putting Algorithmic Bias on Top of the Agenda in the Discussions on Autonomous Weapons Systems," *Digital War* 5 (2024): 201–12, https://doi.org/10.1057/s42984-024-00094-z; Alexander Blanchard and Laura Bruun, *Bias in Military Artificial Intelligence* (Stockholm International Peace Research Institute, 2024), https://doi.org/10.55163/CJFT9557; Ingvild Bode and Ishmael Bhila, "The Problem of Algorithmic Bias in AI-Based Military Decision Support Systems," *ICRC Humanitarian Law & Policy Blog*, September 3, 2024, https://blogs.icrc.org/law-and-policy/2024/09/03/the-problem-of-algorithmic-bias-in-ai-based-military-decision-support-systems/; Arthur Holland Michel, *Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making* (International Committee of the Red Cross, 2024), https://shop.icrc.org/decisions-decisions-decisions-computation-and-artificial-intelligence-in-military-decision-making-pdf-en.html; Jonathan Kwik, "Digital Yes-Men: How to Deal With Sycophantic Military AI?," *Global Policy* 16, no. 3 (2025): 467–73, https://doi.org/10.1111/1758-5899.70042.

55 Ingvild Bode and Tom F.A. Watts, *Meaning-Less Human Control: Lessons from Air Defence Systems on Meaningful Human Control for the Debate on AWS* (Drone Wars UK & Center for War Studies, 2021), https://dronewars.net/2021/02/19/meaning-less-human-control-lessons-from-air-defence-systems-for-lethal-autonomous-weapons/; Ingvild Bode and Tom F.A. Watts, *Loitering Munitions and Unpredictability: Autonomy in Weapon Systems and Challenges to Human Control* (Center for War Studies & Royal Holloway Centre for International Security, 2023), https://www.autonorms.eu/loitering-munitions-and-unpredictability-autonomy-in-weapon-systems-and-challenges-to-human-control/; Anna Nadibaidze, Ingvild Bode, and Qiaochu Zhang, *AI in Military Decision Support Systems: A Review of Developments and Debates* (Center for War Studies, 2024), https://www.autonorms.eu/ai-in-military-decision-support-systems-a-review-of-developments-and-debates.

56 Holland Michel, *Decisions, Decisions, Decisions*; Kwik, "Digital Yes-Men."

57 On LLMs in international security, see Ioana Puscas, *Large Language Models and International Security: A Primer* (United Nations Institute for Disarmament Research, 2024), https://unidir.org/publication/large-language-models-and-international-security-a-primer/.

AI systems cannot be designed and used without meaningful human agency as a safeguard that contributes to mitigating such risks.

Sixth, planning for practices that sustain and strengthen the exercise of human agency addresses the phenomenon of nominal human oversight by providing safeguards against cognitive dependence and skill atrophy. Simply involving humans in use-of-force decision-making at the stage of operational use has proven insufficient to address the multiple concerns surrounding the development and use of AI systems in the military domain, particularly in targeting. There is a risk that only involving humans at this point means they play only a 'nominal' role in military decision-making.

This nominal role, in turn, creates a risk of human cognitive dependence on AI technologies and, ultimately, of atrophying human decision-making capabilities in a context involving death, destruction, humanitarian implications, and long-term impacts on societies and environments.[58] This is precisely why the human role should involve the exercise of sufficient levels of agency, sustained by practices performed by politicians, strategists, developers, users, lawyers, ethicists, and other actors across the lifecycle of AI systems.[59]

The positive norm of human agency that the AutoPractices project seeks to shape benefits three main audiences. First, it benefits governmental security and defence policymakers as well as, second, military decision-makers, because it can reduce the humanitarian, legal, security, and ethical risks associated with accepting a diminished exercise of human agency in targeting decision-making. Third, it benefits various actors involved in designing and developing military applications of AI because it offers practical guidance on how and where humans have to be involved along the AI lifecycle to ensure the exercise of human agency and to attribute responsibility and accountability to humans.[60]

The three sets of audiences that the AutoPractices tookit aims to reach are global in nature. The positive norm of human agency will therefore mean different things across the three audiences due to their different regional and professional backgrounds. Yet, the practices outlined in the practical toolkit are designed to underline that there are many elements in common.

58 Ingvild Bode in Raluca Csernatoni et al., "Myth, Power, and Agency: Rethinking Artificial Intelligence, Geopolitics and War," *Minds and Machines* 35, no. 37 (2025), https://doi.org/10.1007/s11023-025-09741-0.

59 AutoPractices' argument for strengthening the exercise of human agency is based on the assumption that the various human decision-makers involved throughout the lifecycle framework principally act in good faith and not with the explicit intention of violating legal, ethical, or security principles.

60 We recognise that there are important legal distinctions to be made between responsibility and accountability. At the same time, these notions are often used interchangeably in the international debate about AI in the military domain. Stakeholders remarked that the distinctions found in English do not translate in straightforward ways to other languages such as French, German, or Mandarin. See Ariel Conn and Ingvild Bode, "Establishing Human Responsibility and Accountability at Early Stages of the Lifecycle for AI-Based Defence Systems," *Ethics and Information Technology* 27, no. 51 (2025), https://doi.org/10.1007/s10676-025-09862-1.

B Why we need a bottom-up process for governing the development and use of military AI systems

The growing integration of autonomous and AI technologies into targeting decision-making is well-documented and has triggered substantial international debate about associated governance demands for more than 10 years.[61] This debate started with a focus on AWS under the auspices of United Nations Convention on Certain Conventional Weapon Systems (UN CCW) and has since diversified across UN forums, such as the UN General Assembly First Committee, and multi-stakeholder initiatives, such as the Responsible AI in the Military Domain (REAIM) summit series.

States continue to disagree about the necessity of negotiating new, legally binding regulation, but the international debate has produced lists of top-down, 'soft law', high-level principles such as the 2019 Guiding Principles by the Group of Governmental Experts on emerging technologies in the area of Lethal Autonomous Weapon Systems (GGE on LAWS), the 2023 REAIM Call to Action, and the 2024 REAIM Blueprint for Action.[62] Such lists of principles are significant because they indicate an emerging international consensus, including, notably, of the need to retain human control over use-of-force decision-making.[63]

Their high-level character, however, also means that such principles do not spell out a path towards potential implementation, for example around what needs to be done to achieve human control in targeting decision-making involving AI technologies. This is an important point for two reasons: first, more detailed top-down provisions that would ease the practical implementation of high-level governance principles such as human control cannot be expected soon.[64] The governance debate has been moving forward slowly and has increasingly become embroiled in a charged geopolitical atmosphere.

Second, simultaneous to the growing debate about AI in the military domain, AI technologies are already being used in targeting decision-making and have therein already shaped social norms, defined as understandings of appropriateness. This has been the main, thoroughly documented finding of the ERC-funded AutoNorms project.[65]

Such practical processes are therefore also already shaping what should count as a requisite quality of human control. These practical processes appear to currently lead towards accepting a compromised, diminished quality of human control. This is due to the complexity of the tasks military personnel need to perform and the demands they are placed under, for example in terms of speed and interacting with multiple, networked AI-based systems.

61 Ingvild Bode et al., "Prospects for the Global Governance of Autonomous Weapons: Comparing Chinese, Russian, and US Practices," *Ethics and Information Technology* 25, no. 5 (2023), https://doi.org/10.1007/s10676-023-09678-x; Alexander Blanchard and Netta Goussac, *Towards Multilateral Policy on Autonomous Weapon Systems* (Stockholm International Peace Research Institute, 2025), https://doi.org/10.55163/VLLZ7787.

62 UNODA, "Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems," September 25, 2019, https://documents.unoda.org/wp-content/uploads/2020/09/CCW_GGE.1_2019_3_E.pdf; Government of the Netherlands, "REAIM 2023 Call to Action," February 16, 2023, https://www.government.nl/documents/publications/2023/02/16/reaim-2023-call-to-action; Ministry of Foreign Affairs Republic of Korea, "REAIM Blueprint for Action," September 11, 2024, https://www.mofa.go.kr/www/brd/m_4080/down.do?brd_id=235&seq=375378&data_tp=A&file_seq=9.

63 Bode, *Emerging Norms around Military Applications of AI*.

64 The AutoNorms Project, "Global Governance of AI in the Military Domain," September 28, 2023, https://www.autonorms.eu/global-governance-of-ai-in-the-military-domain/.

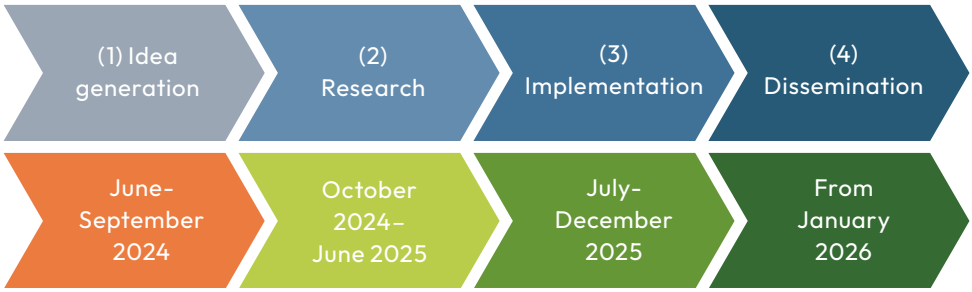
65 Ingvild Bode, "Practice-Based and Public-Deliberative Normativity: Retaining Human Control over the Use of Force," *European Journal of International Relations* 29, no. 4 (2023): 990–1016, https://doi.org/10.1177/13540661231163392; Bode and Watts, *Meaning-Less Human Control*; Bode and Watts, *Loitering Munitions and Unpredictability*; Nadibaidze, Bode, and Zhang, *AI in Military Decision Support Systems*.

This dual finding has inspired the AutoPractices project to sustain a positive norm of human control/agency through a bottom-up process. Such a process entails adapting and reinforcing best practices of designing, training personnel for, and using systems integrating AI technologies in targeting decision-making. This is a bottom-up approach to governing AI technologies in the military domain because it starts from how stakeholders in technical, military, and political spaces perform their (daily) practices. This process can re-shape the emerging norm of human control/agency into a positive version through stakeholders enacting best practices from the bottom up.

C The AutoPractices process of social innovation

The AutoPractices project followed a four-stage pathway to social innovation, that is towards shaping a positive norm of human control/agency in targeting decision-making involving AI technologies through stakeholders' practices from the bottom up. In the following, we outline the four stages of the AutoPractices project: (1) idea generation, (2) research, (3) implementation, and (4) dissemination. Figure 1 offers an overview of these stages and their respective timelines.

Figure 1 AutoPractices project: stages and timeline



Stage 1: Idea generation

AutoPractices starts from AutoNorms' main findings that practices of use are potential sources of both negative and positive emerging norms on the exercise of human agency. This means that practices stakeholders perform throughout the lifecycle of systems integrating AI can diminish, but also sustain and strengthen, human agency in warfare. This basic argument guided the first step of idea generation.

Stage 1 began with the AutoPractices team conducting extensive literature reviews to define key concepts in the project, namely the exercise of human agency and the lifecycle of AI systems.

Human agency

For the purposes of the AutoPractices project, the exercise of human agency in the context of interacting with AI systems is defined as:

> *The capacity to 1) understand and reasonably foresee a system's functions and effects in a relevant context; and 2) to deliberate and decide upon suitable actions in a timely manner; and 3) to act in a way that can impact the use of the system.*

This definition is based on a literature review conducted by the AutoPractices team, which revealed the following common elements in the exercise of human agency:

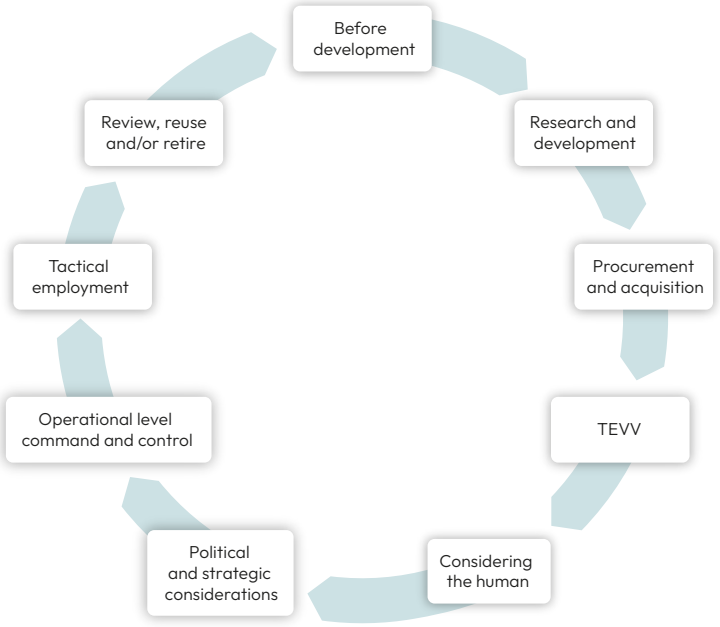
- knowledge of the context (situational awareness) and of the system (its technical characteristics, capabilities, limitations),
- based on this knowledge, the ability to foresee how the system would function and its potential effects in the context of use,
- the ability to reflect/deliberate upon the effects of this system's use, and following this deliberation, the ability to decide on an action/response and act upon this response (the ability to intervene in a timely manner), and the ability to engage in actions that make an impact/change in the world.

The lifecycle of AI systems

The AutoPractices project takes as a foundation the model of the AI lifecycle proposed by the IEEE Standards Association Research Group on Issues of Autonomy and AI for Defense Systems (figure 2). The framework presents a granular way of thinking about the lifecycle with multiple points of human involvement and opportunities to exercise agency. In this model, the lifecycle of a military system integrating AI technologies includes the following 9 stages, which do not necessarily occur in a linear fashion:

1 Before system development
2 Research and development
3 Procurement and acquisition
4 Testing, evaluation, verification, and validation (TEVV)
5 Considering the human: Education, training, and human-system integration
6 Political and strategic considerations
7 Operational-level command and control
8 Tactical employment
9 Review, reuse and/or retire.

Figure 2 Lifecycle framework for systems integrating AI and autonomous technologies in the military domain. [66]



Moreover, the IEEE Research Group highlights five activities that are ongoing across the 9 stages of the lifecycle:

1. evaluation of legal, ethical, and policy concerns
2. responsibility, accountability, and knowledge transfers
3. considering the human: training, education and human–system integration
4. TEVV, monitoring, hardware system or software updates and interoperability, maintenance
5. risk assessments.

Stage 2: Research

This stage mapped practices through engaging with diverse stakeholders. A total of 49 stakeholders took part in the AutoPractices project, either by completing the online survey questionnaire, being interviewed by a member of the AutoPractices team (in person or online) or participating in one of the two AutoPractices workshops (in person or online).

The AutoPractices stakeholders represent five main groups: (1) political–ethical, (2) technical, (3) legal, (4) military, and (5) civil society experts. They were selected based on their knowledge and expertise in relation to AI in security and defence, as well as AI technologies more broadly. All outputs of the AutoPractices project attempt to represent stakeholders' perspectives. This includes using terms employed by stakeholders in survey responses, interviews, and workshops, as closely as possible.

Stakeholders who gave explicit consent to be named in this toolkit are listed in the acknowledgements section. If no such consent was given, the stakeholders remain anonymous.

66 IEEE SA Research Group on Issues of Autonomy and AI in Defense Systems, *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications* (IEEE SA, 2024), https://ieeexplore.ieee.org/document/10707139.

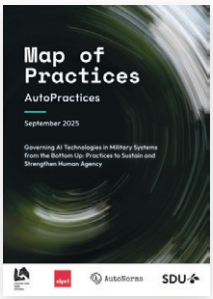
Survey questionnaire

The AutoPractices team designed the survey questionnaire with the intention of using broad, open-ended questions to allow stakeholders to express and reflect upon practices they consider influential to sustain the exercise of human agency across the lifecycle of AI systems. The survey was administered via the SurveyMonkey platform. The link was shared with stakeholders via email. All questions were open-ended (with no limitations on numbers of characters or words) and optional. They appeared in the same order for all respondents. Stakeholders also had the option to leave their name (if they consented to a follow-up interview or workshop) or not. Surveys were shared with various stakeholders from October 2024 to March 2025.

Interviews

Interviews were conducted by members of the AutoPractices team either in person or online and recorded via Microsoft Teams. All transcripts were double-checked by the AutoPractices team and anonymised. The interviews followed a semi-structured form and there were no set questions applying to all interviewees. Questions were therefore adapted to different stakeholders depending on their area of expertise.

Output 1: Map of Practices



Based on an analysis of stakeholders' survey and interview responses, the AutoPractices team produced a **first output**, the *Map of Practices*. This document brings together practices named by stakeholders and organises these into eight key themes. The *Map of Practices* is a transitional, working document upon which this toolkit builds. This document also served as the basis for discussion at the two AutoPractices stakeholder workshops.[67]

Workshops

The AutoPractices team held two workshops with stakeholders: one online (on 26 May 2025) and one in-person held at the University of Southern Denmark in Odense (on 17 June 2025). The workshops brought together mixed groups of stakeholders with different backgrounds to reflectively exchange knowledge about different practices to sustain the exercise of human agency.

The workshops were organised into two main parts. First, stakeholders discussed the draft map of practices in smaller groups and then the whole group. Second, stakeholders engaged in a scenario exercise designed to discuss different challenges/trade-offs of human-AI interaction along selected stages of the lifecycle of systems integrating AI. The scenario exercise was designed to get stakeholders to reflect further about concrete practices that would strengthen the exercise of human agency when facing such challenges/tradeoffs.

This scenario was initially prepared for and presented at the session "Responsible Human-Machine Teaming Across the AI Lifecycle: An Interactive Scenario" co-organised by the AutoNorms project team together with the Johns Hopkins University Applied Physics Laboratory (JHU/APL) at the Responsible AI in the Military Domain (REAIM) Summit which took place in Seoul, Republic of Korea, on 9-10 September 2024.[68]

67 The AutoPractices Project, *Map of Practices* (Center for War Studies, 2025), https://www.autonorms.eu/governing-ai-technologies-in-the-military-domain-from-the-bottom-up-map-of-practices/.

68 For a summary of the session see Qiaochu Zhang, "Navigating the Complexities of Exercising Human Agency in Human-Machine Interaction Across the AI Lifecycle," *The AutoNorms Blog*, October 21, 2024, https://www.autonorms.eu/navigating-the-complexities-of-exercising-human-agency-in-human-machine-interaction-across-the-ai-lifecycle/. A recording of the session is available at: https://www.youtube.com/watch?v=yk2XvDKaChE.

Stage 3: Implementation

Following the mapping of practices and in-depth discussion of this mapping with stakeholders, the AutoPractices team co-created the **second and main output**, the present toolkit. As described, the entire research process of the AutoPractices was designed in close collaboration with its stakeholders.

The toolkit itself was therefore produced iteratively in three steps: first, members of the AutoPractices team drafted the initial version of the toolkit based on the information gathered in the surveys, interviews, and workshops. As part of this process, we referred back to the *Map of Practices* but also went beyond it through closely examining all data provided by stakeholders. Second, we shared the initial version of the toolkit with the 21 stakeholders who participated in the two workshops for comments and feedback. Third, the AutoPractices team adjusted the toolkit based on this stakeholder feedback, thereby finalising it.

Stage 4: Dissemination

The AutoPractices project will spread the best practices toolkit among stakeholders and other key actors in the debate on AI in the military domain. In this way, the toolkit can become a source for sustaining and diffusing a positive norm of human agency in systems integrating AI technologies from the bottom up.

In the short-term, stakeholders, as co-creators of the toolkit, contribute to raising awareness of how different practices shape an emerging norm on human agency in the military domain.

In the long-term, actors adjusting their practices throughout the lifecycle of systems integrating AI contributes to sustaining and strengthening the exercise of human agency. This will likely be a gradual process: new practices cannot replace old ones immediately. The performance of new practices will depend on the extent to which existing policies are in place that allow the necessary leeway for their implementation.

Shared conclusions across the stakeholders indicate a consensus for shaping a positive norm of human agency in the military domain around these practices. We therefore see momentum towards initiating a bottom-up process of social innovation which is international and interdisciplinary—and can be thought of as both independent from or complimentary to any potential top-down international governance in this space.

D1 Survey questionnaire

Dear Participant,

Thank you for agreeing to take part in this survey. The results of this questionnaire will feed into the project "Governing AI Technologies in Military Systems from the Bottom Up: Practices to Sustain and Strengthen Human Agency" (AutoPractices), funded by the European Research Council (Proof of Concept grant no. 101156237).

The purpose of the AutoPractices project is to initiate and accompany a process of social innovation to govern autonomous and AI technologies in the military domain. The project will co-create a set of best practices with stakeholders in the form of a practical toolkit to sustain and strengthen human agency and accountability for the use of AI systems in the military. To co-create this practical toolkit, we approach stakeholders across diverse professional backgrounds and geographies. You have been selected as a research participant based on your knowledge and expertise on AI systems, including in the military domain.

Participation is voluntary, and you are free to decline to answer some questions and can leave the data collection activity at any time without giving a reason. The research team will keep your data strictly confidential and anonymised.

This questionnaire focuses on your understanding (based on your respective area of expertise) of how human agency should be exercised throughout the lifecycle of systems integrating AI technologies.

For the purposes of this survey, the exercise of human agency in the context of interacting with an AI system is defined as: The capacity to 1) understand and reasonably foresee a system's functions and effects in a relevant context; and 2) to deliberate and decide upon suitable actions in a timely manner; and 3) to act in a way that can impact the use of the system.

The lifecycle of a military system integrating AI includes the following stages: 1) before AI system development; 2) research & development; 3) procurement & acquisition; 4) Test, Evaluation, Validation and Verification (TEVV); 5) education & training; 6) political and strategic considerations; 7) operational level command and control; 8) tactical employment; and 9) review, reuse and/or retire.

Questions

Q1. What do you think ensuring the exercise of human agency across the lifecycle of a military system integrating AI means? Please respond in max. 1–2 sentences.

Q2. Based on your area of expertise, what are the most important concerns to address at various stages of a system's lifecycle (listed above) to ensure the exercise of human agency? Please name your top 3 concerns.

Q3. What do you consider as key activities contributing to ensuring the exercise of human agency when it comes to human-system integration/human-machine interaction throughout an AI system's lifecycle? Please name your top 3 activities.

Q4. What types of activities would be detrimental to the exercise of human agency across an AI system's lifecycle? Please name your top 3 activities that you find detrimental.

Q5. What do you consider key activities to ensure responsibility and accountability throughout an AI system's lifecycle? Please name your top 3 activities.

Q6. What are some of the key activities needed to ensure the exercise of human agency in Testing, Evaluation, Validation and Verification (TEVV), monitoring, and maintenance? Please name your top 3 activities.

Q7. Based on your area of expertise, do you consider certain training and education measures essential to ensure the exercise of human agency across a system's lifecycle? If yes, which ones? Please name a maximum of 3 measures.

Q8. How should risks related to systems integrating AI technologies be assessed? Are there any existing frameworks which you would consider useful or important?

D2 Scenario used during the workshops

This scenario was prepared by Ariel Conn and Ingvild Bode, with additional feedback from Lynn Piaskowski, for the breakout session "Responsible Human-Machine Teaming Across the AI Lifecycle: An Interactive Scenario", co-organised by the AutoNorms project and Johns Hopkins University Applied Physics Laboratory (JHU APL) at the Responsible AI in the Military Domain (REAIM) Summit 2024 in Seoul.[69]

Stage 1: Before development

Storyline: As geopolitical tensions rise, military leadership proposes developing new software that can help an uncrewed underwater vehicle (UUV) swarm send more information to operators to enable operators to make more informed decisions about potential targets. The UUV swarm will include AI-based decision-making, specifically for tracking, target identification, and classification, which can assist in potential engagement decisions. This choice raises questions about the role of these operators in the decision-making process. Please choose between two options:

- Choice A: The UUV swarm feeds real-time data to human operators, but the UUVs decide which data is sent and prioritized on the operators' interfaces. The UUV swarm is empowered to make more decisions.
- Choice B: The UUV always sends all information to the operator. Due to network bandwidth limitations, sending this much data can result in latencies. The UUV swarm is empowered to make fewer decisions.

After the audience has chosen between the two options, the following open question will be asked:

- What are the factors that are most important in these trade-offs? Please respond with 2-3 key words.

Stage 2: Research & Development

Storyline: At the R&D stage, the researchers decide to follow a system of documenting and assuring human responsibility throughout the lifecycle. They decide upon a choice between two options for realizing this:

- Choice A: The researchers develop a rigorous system of documenting responsibility and accountability, clearly defining roles and decision-making authority for all members of the development team as well as the operators and the UUVs once the swarm is live. All design choices, test results, and operational decisions are clearly documented and attributed. At crucial handover points, the system integrates accountability checkpoints requiring sign-offs by key stakeholders. This choice slows down the development process significantly.
- Choice B: The researchers develop a flexible documentation strategy that prioritises high-level decision-making processes of the project. This system does not prioritise comprehensive detail but allows individual team leaders in the development process to determine the level of documentation required for the processes related to their specific area of responsibility. This choice includes periodic reviews and results in a more rapid process for system design and development.

69 A recording of the breakout session is available at: https://www.youtube.com/watch?v=yk2XvDKaChE.

Stage 3: Test & Evaluation, Verification, Validation (TEVV)

Storyline: The UUV swarm is now available in prototype form and goes through TEVV. To determine an acceptable error/uncertainty rate for the swarm to exhibit regarding object tracking and identification/classification, researchers develop a computer model that can simulate the UUV swarm's operations by running the simulations thousands of times with different inputs driving these error and uncertainty rates. The model also assesses how the UUV will interact with human operators across various, context-diverse scenarios, thereby assessing their impact on mission success and identifying potential failures resulting from their use. Researchers get feedback on these results from a panel of experts including military strategists, roboticists, ethicists, and legal advisors. The system is designed such that, if one of the UUVs comes across an object it cannot identify, it notifies the operator for guidance on how to proceed. Thereafter, the researchers choose between two options:

- Choice A: If the operator recognises the object (and it is not a threat), the operator can update the swarm logic to recognise this object as a non-threat in all situations moving forward, and thus, the UUVs will no longer check with the operator regarding this object. This will cause/allow error/uncertainty rates of the swarm to change in real-time during operations.
- Choice B: Although the operator can use the interface to log this specific object as one that is non-threatening, the operator cannot update the swarm logic in real-time. Instead, the systems must go through a TEVV cycle to model logic changes and all error/uncertainty rates remain static during real-time operations.

Stage 4: Considering the human: Education, training, and human-system integration

Storyline: As the UUV swarm prototype is drawing closer to potential deployment, researchers consider different choices of designing its user interface. Please choose just one of these three options:

- Choice A: Only high-level information is provided to the user interface with no way for operators to access more detailed/lower-level information. This is to streamline the user interface and prevent overload/confusion for the operators.
- Choice B: Information is provided at varying levels; the default user interface provides high-level information only, but an operator can drill down into all the information gathered by the UUVs if they so desire. The user interface cannot be customized and is always the same, regardless of operator or scenario.
- Choice C: The user interface is dynamic. At times, it presents only high-level information to the operator, at other times, it provides detailed information depending on whether it believes the operator may need that information to make a decision or whether that information is particularly important.

After the audience has chosen between the two options, the following open question will be asked:

- What are the factors that are most important in these trade-offs? Please respond with 2-3 key words.

Final discussion

- What is your main take away from the scenario when it comes to the exercise of human agency?
- In your view, which options would enable the exercise of human agency at various stages?
- Which practices in the scenario do you consider particularly important? Or which ones are missing?

About the AutoPractices Team

Dr Ingvild Bode is a Professor of International Politics and Director of the Center for War Studies at the University of Southern Denmark (SDU). She is the Principal Investigator of the AutoPractices and AutoNorms projects. Professor Bode's research examines how applications of AI in the military domain change international norms, especially on the use of force. Her research has been published in journals such as *European Journal of International Relations*, *European Journal of Public Policy*, *Global Studies Quarterly*, and *Review of International Studies*.

Dr Anna Nadibaidze is a Postdoctoral Researcher at the Center for War Studies, University of Southern Denmark (SDU), and a researcher for the AutoPractices and AutoNorms projects. Her research examines military applications of AI and the global governance of AI in the military domain. Her work has been published in journals such as *Contemporary Security Policy*, *Journal of International Relations and Development*, and *European Journal of International Security*.

Dr Alexander Blanchard is a Senior Researcher in the Governance of Artificial Intelligence (AI) Programme at the Stockholm International Peace Research Institute (SIPRI). His work focuses on issues related to the development, use and control of military applications of AI. He was previously the Defence Science Technology Laboratory (Dstl) Digital Ethics Fellow at the Alan Turing Institute, London.

Shimona Mohan is an Associate Researcher with the Gender & Disarmament and Security & Technology programmes at the United Nations Institute for Disarmament Research (UNIDIR). She was listed as one of the 100 Brilliant Women in AI Ethics for 2024, and holds a Master in International Affairs from the Geneva Graduate Institute in Switzerland. She contributes to AutoPractices in her personal capacity.

Ariel Conn is an Instructor in Computing Ethics at the University of Colorado, Boulder, and she is the founder and President of Mag10 Consulting, where she works with researchers and not-for-profit organizations to address a variety of global ethical and policy issues. Her work is primarily focused on the use of AI in the military and other ethically challenging uses and impacts of AI. She is an expert advisory member for the Global Commission on Responsible AI in the Military Domain, and she was one of 2023's 100 Brilliant Women in AI Ethics.

Dr Hendrik Huelss is an Associate Professor of International Relations at the Center for War Studies, University of Southern Denmark and an associated senior researcher in the AutoNorms project. His research combines an interest in norms in International Relations with critical perspectives on technologies in politics. He has published in journals such as *European Journal of International Security*, *International Political Sociology*, *International Theory*, and *Review of International Studies.*

Dr Qiaochu Zhang is a Max Weber Postdoctoral Fellow at the Florence School of Transnational Governance, European University Institute. Her research focuses on global AI governance and Chinese foreign policy. Her work has been published in *International Affairs*, *International Peacekeeping*, and *Global Policy: Next Generation*, among others.

Dr Guangyu Qiao-Franco is Assistant Professor of International Relations at Radboud University and Advisory Board Member at the Leiden Asia Centre. She is Principal Investigator of a Dutch MOFA-funded project on China's export control strategy. Her research focuses on AI governance, export controls, and Global South perspectives in international politics.

Dr. Tom F.A. Watts is a Postdoctoral Researcher at the Center for War Studies, University of Southern Denmark. Previously, he was a Leverhulme Early Career Fellow based at Royal Holloway, University of London. His research examines the relationship between great power competition and military applications of AI. His work has been published in leading International Relations journals including *Cooperation and Conflict, Geopolitics,* and *International Politics*, among others.


CENTER FOR
WAR
STUDIES


sipri


AutoNorms


SDU
University of
Southern Denmark