UNIDIR

Draft Guidelines for the Development of a National Strategy on AI in Security and Defence

A Policy Brief

SECURITY AND TECHNOLOGY PROGRAMME

Acknowledgments

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities. This policy brief was prepared by the Artificial Intelligence Workstream of UNIDIR's Security and Technology Programme, which is funded by the Governments of Czechia, France, Germany, Italy, the Netherlands, Norway, the Republic of Korea, Switzerland and the United Kingdom, and by Microsoft.

The author wishes to thank Dr. Giacomo Persi Paoli (UNIDIR) for advice, guidance and support for this programme of work and for his review of this report; as well as Ioana Puscas (UNIDIR) and Dr. Samuel Segun (Global Center on AI Governance) for their reviews and suggestions. The author also wishes to thank states that participated in the focus group discussions for this project in March 2024, as well as members of the Roundtable for AI, Security and Ethics (RAISE) for their valuable contributions and input to the guidelines. Finally, the author wishes to thank Jessica Espinosa Azcárraga, Edward Madziwa and Elia Duran-Smith for their tremendous support in the organization of all the research events for this project.

About UNIDIR

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

Note

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

About the Security and Technology Programme

Contemporary developments in science and technology present new opportunities as well as challenges to international security and disarmament. UNIDIR's Security and Technology Programme seeks to build knowledge and awareness on the international security implications and risks of specific technological innovations and convenes stakeholders to explore ideas and develop new thinking on ways to address them.

Author

This report was produced by UNIDIR's Security and Technology Programme. It was drafted by Yasmin Afina, principal investigator for this project.

Yasmin Afina

Researcher, Security and Technology

Yasmin Afina is a Researcher for the Security and Technology Programme at UNIDIR, where her research covers the intersection between international security, international law and artificial intelligence. Her research experience and interests cover nuclear weapons policy, outer space security, and wider international security and policy issues surrounding emerging technologies, including neurotechnology, quantum technologies, and cyber. She is also a PhD Researcher in law at the University of Essex.

Acronyms & Abbreviations

AI	Artificial intelligence
IHL	International humanitarian law
IHRL	International human rights law
ISR	Intelligence, surveillance and reconnaissance
REAIM	Responsible AI in the Military Domain

Contents

Introduction

Background

As innovation in artificial intelligence (AI) proceeds at breakneck speed, the appetite exhibited by states for devising frameworks for the governance of the research, development and deployment of these technologies is at its greatest. With calls for governance solutions increasing at both the national and international levels, the number of national strategy documents that frame the development, deployment and use of these technologies has started to grow across regions.

Yet, most of these policies exclude or barely touch upon security and defence applications. Only a handful of national strategy documents have a section dedicated to this realm; and even fewer are specifically dedicated to it. This scarcity is at odds with the United Nations Secretary-General's recommendation for Member States to "urgently develop national strategies on responsible design, development and use of artificial intelligence", as outlined in his New Agenda for Peace.[1]

A number of state-led initiatives, such as the Call to Action of the 2023 Responsible AI in the Military domain (REAIM) Summit, have also recognized national strategies as being key to enabling the responsible development, deployment and use of military AI.[2] The importance of the strategies being formulated "to ensure responsible AI applications in the military domain" has been further stressed in the newly adopted REAIM Blueprint for Action.[3]

Against this backdrop, UNIDIR has launched a programme of work to establish guidelines for the development, adoption, implementation and review of national strategies on AI in security and defence. Over the next few years, this programme of work will capture, anticipate and dissect the key issues, considerations and needs that must be addressed as states grapple with the issue of responsible AI in security and defence. It will do this by drawing on states' good practice and their shared methods in this space, along with input from non-state stakeholders with a role in the AI governance ecosystem.

Purpose and Structure of the Guidelines

The purpose of the guidelines is to capture, anticipate and dissect the key issues, considerations and needs that each state must address as it develops or seeks to develop, adopt, implement and review its national strategy on AI in security and defence. In recognition of the host of incentives stemming from the establishment of such strategies, as described in further detail in the following section, it is hoped that these guidelines will serve as a useful tool for states and non-state stakeholders alike as they seek to address issues related to the responsible development, deployment and use of AI in security and defence.

---

[1] United Nations, *A New Agenda for Peace*, Our Common Agenda Policy Brief 9 (New York: United Nations, July 2023), https://dppa.un.org/en/a-new-agenda-for-peace, p. 28.

[2] Responsible AI in the Military domain (REAIM) Summit, "REAIM Call to Action", 15–16 February 2023, https://www.government.nl/documents/publications/2023/02/16/reaim-2023-call-to-action, para. 5.

[3] Responsible AI in the Military domain (REAIM) Summit, "REAIM Blueprint for Action", 9–10 September 2024, https://www.mofa.go.kr/www/brd/m_4080/down.do?brd_id=235&seq=375378&data_tp=A&file_seq=, para. 8.

While they do not seek to be prescriptive, the guidelines will provide a series of considerations and recommendations divided into two types:

1. Procedural guidelines

Part I of the guidelines is dedicated to providing states with an anthology of recommendations on the process surrounding the development, adoption, implementation and review of national strategies on AI in security and defence. These guidelines building on good practices, meaningful efforts and existing approaches adopted and considered by states at the national, regional and international levels. It is hoped that the procedural guidelines will provide states and non-state stakeholders with food-for-thought on key steps, processes and considerations to support the development, adoption, implementation and review of a national strategy on AI in security and defence.

2. Substantive guidelines

Part II of the guidelines is dedicated to outlining a series of substantive considerations that states should examine and integrate, as appropriate, in the development, adoption, implementation and review of a national strategy on AI in security and defence. While in no way exhaustive of all relevant considerations – which are highly context-dependent, varying from one state to another and across regions – it is hoped that the substantive guidelines will provide states and non-state stakeholders with food-for-thought on key issues that may, or may be anticipated to, influence the state's approach and subsequent strategy in this space.

The establishment of these guidelines is further motivated by the recognition that states, both within and across regions, are at different stages in the development, adoption, implementation or review of their national strategies on AI in security and defence.[4] The guidelines will thus serve two purposes: for states in the relatively early stages of developing a strategy, and even more so for those considering doing so, the guidelines will serve as a tool to build and consolidate capacity; for states already at the implementation, or even review, stages of their national strategy on AI in security and defence, the guidelines will remind them of the key policy, legal and ethical considerations surrounding the responsible development, deployment and use of these technologies.

In addition, while the development of a national strategy on AI in security and defence would require the mobilization of considerable resources and time, it is hoped that these guidelines will also showcase the dynamic and "living" nature of a strategy. Akin to development of a national strategy in the cybersecurity realm, states are not expected to craft a perfect, all-inclusive and everlasting strategy from the outset. It should rather be viewed as a framework to be developed and refined through implementation, review and updating. It is therefore hoped that these guidelines can encourage states to approach strategy development not as a quest for perfection but as a necessary first step; and that they will serve as a companion for states throughout their journey in developing, adopting, implementing and reviewing their national strategy on AI in security and defence.

---

[4] Yasmin Afina, *The Global Kaleidoscope of Military AI Governance: Decoding the 2024 Regional Consultations on Responsible AI in the Military Domain* (Geneva: UNIDIR, 2024), https://unidir.org/publication/the-global-kaleidoscope-of-military-ai-governance/.

The release of the present draft guidelines aims to provide state and non-state stakeholders alike with an opportunity to share their feedback with UNIDIR. As the guidelines seek to be inclusive of all approaches to the development, adoption, implementation and review of national strategies on AI in security and defence, opening up the draft guidelines for feedback will ensure varying perspectives are incorporated into their final form.

Why a National Strategy on AI in Security and Defence?

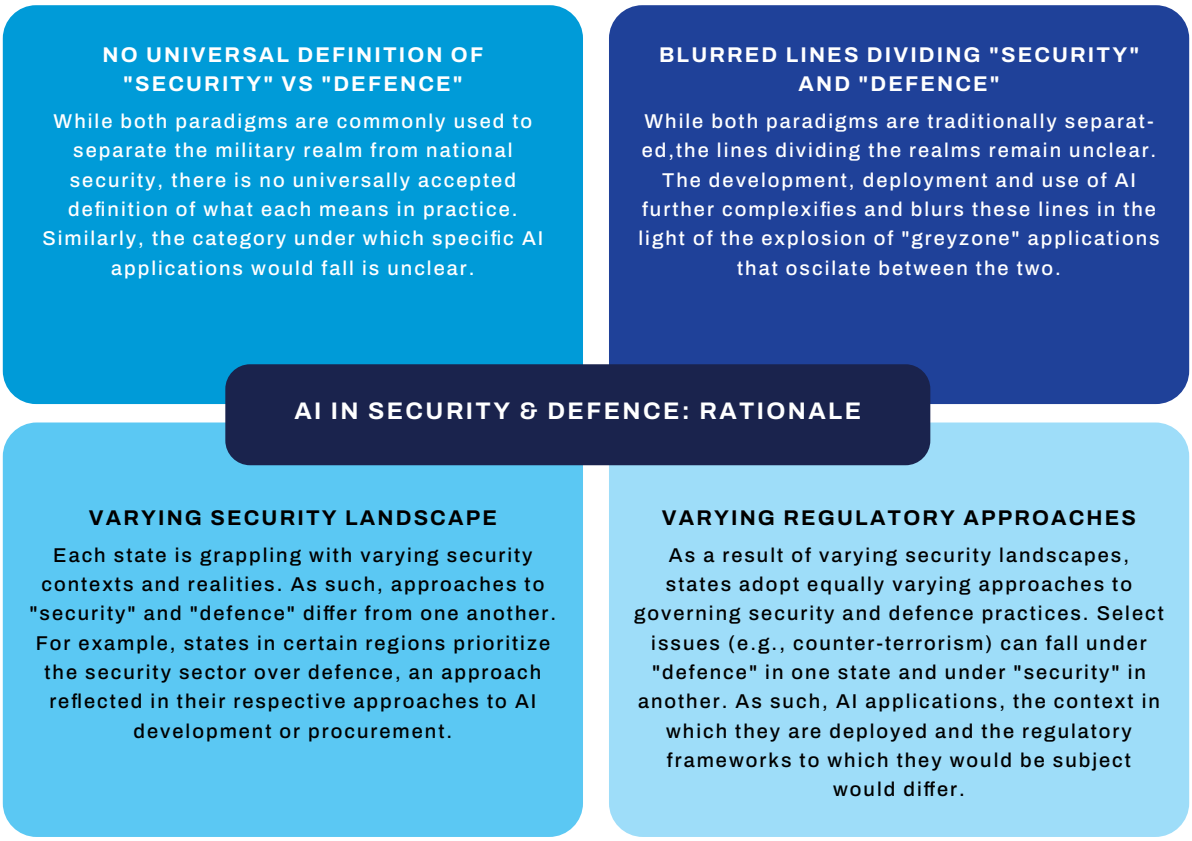
There are a number of reasons and incentives for a state to establish a national strategy on AI in security and defence. While such a strategy can take many forms, its effective development, adoption, implementation and review present many opportunities to build the state's capacity, readiness and maturity in the face of technological progress in this space:

- **Clear state-of-the-art:** A comprehensive survey and understanding of the technological landscape in security and defence are essential for the development, adoption, implementation and review of a national strategy. This will not only provide the state with clarity on the state-of-the-art with regards to technological progress; this will also enable it to comprehensively map relevant stakeholders as well as their respective roles and responsibilities.
- **Evident distribution of roles and responsibilities:** Building on the comprehensive mapping of relevant stakeholders in the governance of AI in security and defence, a national strategy will provide clarity on the distribution of roles and responsibilities and, when appropriate, the working relationships between actors, agencies and organizations, both public and private.
- **Effective coordination:** In addition to a clear distribution of roles and responsibilities, a national strategy also provides a framework to enable effective coordination between actors, agencies and organizations, both public and private. Beyond communication channels, effective coordination is necessary to ensure alignment in implementing existing policies, laws and ethical guidelines, thus preventing fragmentation between government agencies, organizations and between the public and private sectors. Effective coordination will also be needed to optimize the necessary financial, human and technological resources.
- **Change management:** In recognition of AI's general-purpose and cross-cutting nature, a strategy will provide a clear and stable framework to manage the organizational changes that the development, deployment and use of these technologies will entail in security and defence.
- **Resource mobilization:** Setting a state's technological and regulatory objectives, aspirations and vision in a national strategy will enable the mobilization and subsequent allocation of the financial, human and technological resources required to operationalize the state's approach to the governance of AI in security and defence.
- **Public oversight:** A national strategy on AI in security and defence will demystify the possible applications of these technologies and the frameworks within which they are developed, deployed and operate. The general public will be informed of the state's intention, plans and approaches surrounding these technologies, thus enabling appropriate public oversight.

- **Responsible innovation:** The development, deployment and use of AI in security and defence present a host of opportunities, from enhanced intelligence, surveillance and reconnaissance (ISR) capabilities to logistics support and management. As public and private actors alike seek to harness these opportunities, strategy documents act as a means to ensure a responsible approach to AI innovation. National strategies also serve as a means to formally mainstream specific aspects of "responsible AI", such as ethics, bias mitigation and data governance.
- **Risk reduction:** In addition to the opportunities offered by AI technologies in security and defence, the risks stemming from their development, deployment and use must be addressed. By capturing the risks that these technologies present, a national strategy offers an opportunity for a state to identify risk-reduction and -mitigation approaches, and a concrete plan for their subsequent implementation and operationalization. Specific considerations for security and defence applications to take account of the dual-use nature of AI technologies will be particularly important, as innovation in the civilian space proceeds at breakneck speed.
- **International collaboration:** A national strategy provides clarity for the international community on a state's objectives, state-of-the-art and overall approach to the governance of AI in security and defence. This clarity not only serves as a confidence-building measure, it will also pave the way for potential collaborations at the interstate, regional and international levels.
- **Interpretation of international law:** The development, deployment and use of AI in security and defence are all framed by a series of international legal frameworks, including international humanitarian law (IHL), international human rights law (IHRL) and *jus ad bellum*. Yet, aside from applicability, there is no universally agreed consensus as to how international law applies. A national strategy offers an opportunity for a state to reflect on how it would approach the interpretation of international law.

Scope

The present draft guidelines seek to cover AI in “security and defence”. The widening of the scope of the guidelines beyond defence (i.e., the military domain) into the broader domain of security is driven by a number of reasons.



Methodology

The draft guidelines are the product of an extensive review of existing national strategies and governmental policy documents pertinent to AI in security and defence, where available in the public domain.[5] Existing guidelines, both related and unrelated to technology security, have also been studied to draw lessons on scoping, structure and delivery.[6]

---

[5] Much of the resources consulted were extracted from UNIDIR’s Artificial Intelligence Policy Portal (AIPP), https://aipolicyportal.org/, a comprehensive repository of policy documents on AI of all United Nations Member States’ and various intergovernmental organizations, and multi-stakeholder and other initiatives.

[6] Among the model guidelines and other documents studied, see in particular International Telecommunications Union (ITU) et al., *Guide to Developing a National Cybersecurity Strategy: Strategic Engagement in Cybersecurity* (Geneva: ITU, 2018), https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-CYB_GUIDE.01-2018-PDF-E.pdf; Noam Lubell, Jelena Pejic and Claire Simmons, *Guidelines on Investigating Violations of International Humanitarian Law: Law, Policy, and Good Practice* (Geneva: Geneva Academy and ICRC, 2019), https://www.icrc.org/sites/default/files/document/file_list/guidelines_on_investigating_violations_of_ihl_final.pdf; Global Task Force for Inclusive AI, “[Draft] Guidelines for Participatory and Inclusive AI”, Partnership on AI, 2024, https://partnershiponai.notion.site/1e8a6131dda045f1ad00054933b0bda0; Centre for Humanitarian Dialogue, “Code of Conduct on Artificial Intelligence in Military Systems”, 2021, https://hdcentre.org/wp-content/uploads/2021/08/AI-Code-of-Conduct.pdf.

In addition, four focus group discussions were held in early 2024 to consult with state representatives from all five regional groupings of United Nations Member States (African States, Asia-Pacific States, Eastern European States, Latin American and Caribbean States, and Western European and other States).[7] A multi-stakeholder consultation with representation from civil society, industry and academia across all five regional groupings also took place in early 2024. Further experts were consulted bilaterally to complement and consolidate UNIDIR's findings.

The present draft guidelines have been released to provide states and all relevant stakeholders involved in the development, adoption, implementation and review of national strategies on AI in security and defence with an opportunity to review and provide feedback to UNIDIR. The Institute aims to adopt a holistic and inclusive method to the establishment of the guidelines; it thus seeks to capture all the varying perspectives, viewpoints and approaches to this issue.

[7] United Nations, "Regional Groups of Member States", n.d., https://www.un.org/dgacm/en/content/regional-groups.

CHAPTER I.

Procedural Guidelines

The first guidelines for the development, adoption, implementation and review of national strategies on AI in security and defence are procedural. They do not seek to be prescriptive but, rather, provide states and key stakeholders with an overview of the main considerations that, albeit non-exhaustive, would benefit from extensive reflections. They also include, as appropriate and necessary, solutions to support the development, adoption, implementation and review of national strategies.

1. Development and Adoption of a National Strategy on AI in Security and Defence

In the development of a national strategy on AI in security and defence, a number of steps and key considerations must be taken into account. These encompass both internal and external factors that may have an impact on the national strategy under development.

1.1. Conceptualization

The development of a national strategy on AI in security and defence requires a clear vision and strategic direction. Hence, a state developing such a strategy should engage in a conceptualization process that would help define and identify its goals in the AI space, and subsequently align those with the broader national security posture. The state should consult extensively with internal and external stakeholders to ensure a comprehensive overview and understanding of the possible opportunities and perceived risks associated with AI in security and defence, and subsequently formulate a clear vision and strategic direction. Each state should consider the establishment of working groups comprising military officers, experts, AI developers, lawyers, ethicists and policymakers from within government and outside. The conduct of regular consultations and working sessions would help define the state's strategic direction and goals.

1.2. Definition of Objectives

A state developing a national strategy on AI in security and defence should define clear objectives that underpin that strategy. These objectives should provide for the state's ambitions and goals, that is, what it ultimately seeks to achieve through the integration of AI in security and defence. To this end, a comprehensive survey of opportunities that may be offered by these technologies and specific applications would be helpful. The state should define specific and measurable objectives, and should create subsequent action plans for each objective. These plans should include a clear description of their scope and underlying rationale, clear timelines, responsible actors, milestones and key performance indicators to track, review and evaluate progress.

1.3. Definition of Desired Outcomes

A state developing a national strategy should also define the desired outcomes of the strategy, reflecting short-, medium- and long-term goals and aspirations in security and defence. To this end, the state should define specific and measurable desired outcomes, considering the national, regional and international security policy landscapes and their interplay. The state should consider the development of monitoring and evaluation frameworks to measure, track, review and evaluate progress towards the desired outcome.

1.4. Definition of Scope

Defining the scope of a national AI strategy in security and defence would be critical to ensuring that the appropriate kinds and levels of resources are allocated and directed towards priority areas, while the strategy remains grounded within a very specific framework. The scope should clearly outline areas where AI would be applied in security and defence, thus ensuring alignment with existing national

security goals. The scope should cover both immediate operational needs, as well as long-term strategic objectives, taking into consideration domestic security concerns, international obligations and technological trends. As such, a state developing a national strategy should consider developing a detailed scoping document that would outline the various applications of AI in security and defence, and limitations and criteria for expanding or narrowing its use over time. This scope should be subject to review on a regular basis, ultimately ensuring that it remains relevant as technology and security environments evolve.

1.5. Distribution of Roles and Responsibilities

One of the most critical aspects of the implementation of an AI strategy in security and defence corresponds to the clear and effective distribution of roles and responsibilities among all stakeholders. Such a distribution would notably address concerns about potential overlaps in roles between military branches, intelligence and law enforcement agencies, as well as ministries, particularly in the light of blurring lines dividing security and defence applications. A state developing a national strategy should consider the development of a framework that codifies such a distribution of roles and responsibilities, which would include specific mandates, powers and limitations for each public body and under which circumstances they apply.

1.6. Internal Consultations

Internal consultations are critical to ensure that all relevant government entities and agencies are aligned on the national AI strategy. These consultations should include, among others, relevant ministries (e.g., ministries of defence, of foreign affairs, of interior, and of information and communications technology), intelligence agencies, law enforcement agencies, and other relevant bodies to ensure a united and unified front and approach. A state developing a national strategy should thus organize and facilitate regular inter-agency meetings to discuss the development, adoption, implementation and review of an AI strategy. This will ensure that the national strategy is holistic and encompasses the varying viewpoints from across the governmental ecosystem. Ensuring effective coordination between agencies will, in this sense, be key and the state should therefore identify a specific agency or body responsible for inter-agency coordination and for facilitating communication.

1.7. External Consultations

Engaging with external stakeholders – including academic and research institutions, civil society organizations, and the private sector – would be critical to building a robust, inclusive and effective national strategy on AI in security and defence. A state developing such a strategy should take stock of the external expertise available, establish a national network of experts, and leverage the latter to address the technical, legal, ethical and operational risks, challenges and concerns raised in this space. The state should consider establishing a formal consultation process with external stakeholders; the consultation would both have a high-level track and collect input, feedback and recommendations on specific issues and topics. The state should then consider the extent to which these external stakeholders should be involved (either formally, informally or both) in the subsequent implementation of the national strategy.

1.8. Assessment of Relevant Geopolitical Dynamics at the International and Regional Levels

A state developing a national strategy on AI in security and defence should integrate assessments of the relevant geopolitical dynamics, issues and concerns at the international and regional levels that may have a direct or indirect impact on the strategy. A thorough

understanding of the international and regional geopolitical environment would indeed be critical for shaping effective AI strategies in security and defence. To this end, the state could consider the establishment of a dedicated working group tasked with the regular monitoring of geopolitical dynamics at the international and regional levels. The state should evaluate the ways in which these dynamics are relevant to and may influence the national strategy under development. This includes, for example, the impact that geopolitical tensions could have on supply chains and subsequent repercussions for strategic technological dependencies. On a macro level, strategic competition and the perceived "AI arms race" to secure military advantage could also have an impact on the state's posture and overall approach to its national strategy (e.g., accelerated procurement processes). As such, the state should consider carefully the implications of international and regional geopolitical dynamics on its national strategy, and possible repercussions on legal compliance and ethics. The extent to which the working group should or should not include external stakeholders, in addition to the specific agencies involved and disciplines represented, is at the discretion of the state in question, although the inclusion of a diversity of perspectives would consolidate the assessments made by the working group.

1.9. Assessment of Relevant Political Dynamics and Context at the National Level

A state developing a national strategy should integrate assessments of the relevant political dynamics at the national level, including public opinion, political stability and legislative environments. The state should assess the political landscape at the domestic level to ensure that the strategy under development will secure internal buy-in. A thorough understanding of the national political landscape would also influence the likelihood of success in implementing the national strategy, and would help to anticipate areas where flexibility may be required in order to adapt to changing conditions. To this end, the state should organize formal and informal briefings and consultations with legislative bodies to take stock of national priorities, gauge political support and address concerns within the national strategy.

1.10. Assessment of the Wider AI Policy and Regulatory Landscape

A comprehensive review of existing AI policies and regulations at the national, regional and international levels is necessary to ensure that a national strategy is cohesive and adequately aligned to the regulatory frameworks in place. To this end, a state developing a national strategy should conduct a comprehensive review and stocktaking exercise of AI policies and regulations to which it is party at the international, regional and national levels or are otherwise relevant. The state should conduct such reviews periodically in order to ensure that the national AI strategy takes into account the rapidly evolving landscape of governance of AI in security and defence, and is adapted accordingly and as necessary.

1.11. Assessment of International Law

The development, deployment and use of AI in security and defence must be conducted in adherence and compliance with international law, including international humanitarian law and international human rights law. As such, a state developing a national strategy should conduct a comprehensive and thorough assessment of applicable legal frameworks and regimes at the international and regional levels. Such assessments should be conducted in consultation with legal experts from various fields, including academia, the armed forces, governmental bodies, and international and regional organizations. This will ultimately ensure that the AI strategy adheres to the state's international and regional obligations.

1.12, Assessment of Ethics

Ethical considerations should lie at the forefront of the development, deployment and use of AI in security and defence. As such, a state developing a national strategy should conduct a comprehensive and thorough assessment of existing ethical guidelines and a survey on the ethical principles of importance. Such an assessment should be conducted in consultation with ethicists from various fields, including academia, governmental bodies, and international and regional organizations. This will ultimately ensure that the AI strategy is aligned with ethical frameworks and guidelines, and with technologies across their life cycle.

1.13. Assessment of Technical Standards

A state developing a national strategy should take stock of and evaluate technical standards in place to ensure that AI systems in security and defence are robust, reliable and interoperable across domains and varying security contexts. These include technical standards adopted by specialized organizations and other bodies (e.g., the International Organization for Standardization (ISO) and the Institute of Electrical and Electronics Engineers Standards Association (IEEE SA)). In addition, the state should consider the development of national technical standards for AI systems in security and defence, in alignment with good practices in this space.

1.14, Assessment of AI Risks

In the development of its national strategy, a state should assess the risks associated with the development, deployment and use of AI in security and defence. These studies should then be complemented by the development and adoption of a clear risk-mitigation framework, with an established risk-assessment process mandated across governmental agencies and throughout the technology's life cycle. The state should also consider the conduct and implementation of regular risk audits, thus ensuring the continued scrutiny of AI technologies in security and defence against existing, new and emerging risks.

1.15. Assessment of Non-AI Security Risks

In addition to AI-specific risks, broader security issues, concerns and threats that may be of relevance to AI in security and defence should be considered. These non-AI security risks include cyber threats, risks posed by the activities of private military contractors and non-state armed groups, as well as geopolitical instability. A state developing a national strategy should thus conduct a comprehensive assessment to identify, map and consider these risks, and ensure that the strategy factors in the broader security landscape and, subsequently, non-AI security risks that may be of relevance. To this end, the state should coordinate with its specialized security bodies and agencies (e.g., national cybersecurity agency, national nuclear authority, intelligence agencies) to take stock on these risks, the extent to which they are of relevance to AI in security and defence, and the role of a national strategy in mitigating such risks.

1.16. Collective Drafting of a National Strategy Document

Building on the studies, consultations processes and assessments undertaken in the first phase of the strategy's development, a state developing a national strategy should consider the establishment of a cross-agency working group mandated to draft a national strategy document. The process should be framed by a specific mandate and clear rules of procedure to ensure that the strategy incorporates diverse perspectives and expertise from security and defence and from across disciplines. The national strategy should be a comprehensive yet flexible framework that would guide the development, deployment and use of AI in security and defence. It should outline

clear objectives, the desired outcomes and governance structures including for its implementation, oversight and review.

1.17. Classification: To Publish or Not Publish?

A state developing a national strategy on AI in security and defence should reflect on and decide whether to publish its strategy. Considerations may include sensitivities and the classification of certain information, the need for transparency, and risks and opportunities that stem from the publication or, conversely, the classification of such a strategy. The state should also consider partial declassification, with a public-facing version of the strategy and a more comprehensive classified version that may contain sensitive information.

1.18. Review of the Draft National Strategy Document

Prior to its adoption, a state developing a national strategy should establish a review process to ensure that key stakeholders are able to provide feedback and input into the text. An institutionalized multistage review process would ensure the strategy's inclusivity and effectiveness. To this end, the state should organize a series of review consultations, workshops and processes to collect and incorporate feedback from both public and private stakeholders.

1.19. Adoption of the National Strategy Document

Once the strategy has been finalized, the state should proceed with its formal adoption and communicate it to all relevant stakeholders. The adoption process may vary from one state to another, depending on the respective regulatory, legislative and policymaking processes. The state should work, in parallel, on a communication plan and strategy to disseminate the national strategy as appropriate to the relevant stakeholders, thus ensuring that they remain informed and prepared to implement their respective duties and responsibilities following the adoption of the strategy.

2. Implementation and Review of a National Strategy on AI in Security and Defence

Following the development and adoption of a national strategy on AI in security and defence, a number of considerations are specifically relevant to its implementation, including the actors involved, monitoring and evaluation processes, the need for oversight and verification mechanisms, and auditing.

2.1. Key Actors for Implementation

The effective implementation of an AI strategy would require the involvement and intervention of various key actors from across government, the military, industry (including military contractors and technology companies), intelligence and security contractors, and academic and research institutions. Each actor plays a unique role in ensuring that the AI strategy is successfully implemented and operationalized. To this end, a state implementing a national strategy should define and codify a clear distribution of roles and responsibilities of key actors, from both the public and the private sectors. A designated lead agency should be identified and mandated to coordinate implementation efforts and facilitate inter-agency communication.

2.2. Key Coordinating Bodies and Actors

A central coordinating body would be critical to enable and oversee the implementation of the national strategy. This coordination body could either be concentrated in a single agency (e.g., embedded within a specific ministry) or could be composed of representatives of diverse agencies. This body should have a clear, explicit and unambiguous mandate, with the authority to manage inter-agency initiatives, efforts and collaborations and to monitor and evaluate the implementation process, as well as the ability to address potential challenges in a swift and timely manner. The coordinating body should have the resources required to regularly operate and organize inter-agency meetings, manage timelines, and monitor implementation milestones. A state implementing a national strategy should also reflect on whom this coordinating body would be accountable and report to, and should establish the appropriate mechanisms in this space.

2.3. Clarity of Structure and Distribution of Roles and Responsibilities

A clear structure for the distribution of roles and responsibilities would be critical for the implementation of the national strategy. The structure should be comprehensive, yet it should maintain a degree of flexibility to adapt to evolving technological and operational needs for the implementation of the strategy. A state implementing a national strategy should thus develop a clear governance structure that clearly delineates roles and responsibilities at the international, regional, national and agency levels. Each entity should be aware of its role and responsibilities, reporting structures, expectations, and ways through which it could contribute to the overall strategy.

2.4. Key Actions for Implementation

A state implementing a national strategy would need to develop a road map of key actions to guide and frame the effective implementation of the strategy. As such, the state should create a detailed action plan that outlines key steps for implementation, including timelines, resource

allocation (including budgeting and funding) and performance metrics. Each key action should then be broken down into projects with clear deliverables, assigned teams and a tracking system to monitor progress.

2.5. Key Timelines and Milestones for Implementation

Establishing realistic timelines and milestones would ensure the effective implementation of the national strategy. These milestones should allow for periodic assessments and reviews as necessary. A state implementing a national strategy should thus develop a timeline for each phase of the strategy, setting specific milestones and frameworks for the tracking, measurement and reporting mechanisms. The state should also consider the need for flexibility as needs be, and adjust the strategy based on technological progress and the ever-evolving policy and regulatory landscape surrounding the governance of AI in security and defence.

2.6. Accountability for Implementation

Accountability mechanisms are necessary to ensure that stakeholders meet and implement their respective roles, duties and responsibilities. These mechanisms should include clear accountability mechanisms that would hold actors responsible for meeting their implementation targets. Possible measures include performance reviews, reporting obligations and providing a coordinating body with an oversight mandate.

2.7. Adoption of an Implementation Plan

Upon the development of a strategic implementation plan, the state should work towards its formal adoption not only at the national level, but also at the agency level and by all stakeholders. By requiring a formal and multi-layered adoption process, such an approach would foster accountability, ownership and commitment towards the implementation of the national strategy.

2.8. Establishment of an Oversight Mechanism

A state implementing a national strategy should consider the establishment of a formal oversight body, tasked with monitoring AI strategy implementation, reviewing progress against milestones, and making the necessary adjustments to keep the implementation of the strategy on track. This body should be conferred with the appropriate mandate and powers needed to undertake its duties, along with the allocation of the required financial, human and technological resources. In its mandate, the state should consider the organization and facilitation of regular inter-agency convenings to assess implementation progress, identify issues and challenges, and adapt accordingly.

2.9. Establishment of a Verification Mechanism

Verification mechanisms would be necessary to monitor, assess and evaluate progress in the implementation of a national strategy. A state implementing such a strategy should thus establish verification protocols with concrete technical solutions including audits, third-party reviews as appropriate, and internal assessments with regards to the implementation of the strategy. Verification mechanisms would also be critical in the monitoring and assessment of AI use in security and defence, particularly in order to establish forensic evidence to feed into formal investigations in cases of incidents that may have an impact, at times severe, on strategy implementation.

2.10. Establishment of a Review Mechanism

A review mechanism would be necessary to conduct periodic assessments of a national strategy's relevance, efficacy and efficiency in the light of novel technological advancements and the evolving security landscape. As such,

a state implementing an AI strategy should consider the implementation of a cyclical review process, with regular assessments scheduled on progress, on relevance and against the wider security landscape. The establishment of a review mechanism and process would be instrumental for the state to lay the groundwork for the review of the strategy as a whole and, eventually, for the development and adoption of updated editions of the strategy.

3. Cooperation

Internal and external cooperation will play a critical role in the development, adoption, implementation and review of a national strategy on AI in security and defence. As such, states should consider how they intend to frame such efforts in the light of the sensitive nature of the topic at hand.

3.1. National Cooperation

National cooperation between government agencies and branches of the armed forces would be crucial for ensuring the successful and effective implementation of a national strategy on AI in security and defence. To this end, a state implementing such a strategy should introduce a national cooperation framework that would codify the respective mandates of each agency, and possible working relationships for the implementation of the strategy. The state should also consider the establishment of inter-agency working groups with a thematic focus, enabling collaboration, coordination and alignment on specific security and defence issues.

3.2. Bilateral Cooperation

Bilateral cooperation with other states would be crucial for the sharing of good practices, for the alignment of policies and regulations, as well as for enhancing security and defence relationships (e.g., through joint investment programmes and initiatives). The development of bilateral AI cooperation agreements with key allies may be particularly conducive due to the limited scope of such agreements and considering the sensitivities surrounding the development, deployment and use of AI in security and defence. A state implementing a national strategy should consider the degree and depth of such bilateral cooperation, the relevant actors, and the resources that would need to be allocated, for what specific purposes and within what timeframes.

3.3. Regional Cooperation

Regional cooperation, where desirable and feasible, would allow a state to align its AI strategy with those of other states in the region and with regional security frameworks. Such cooperation and partnerships could pave the way for collaboration to address common challenges, reinforce regional security and stability, and ensure technical interoperability between systems. As such, states should promote regional cooperation frameworks through the organization of convenings and summits specific to AI in security and defence, both high-level and at the working level. Joint regional defence projects with regards to AI development and integration in security and defence should also be considered, notably to address challenges and issues that affect the regional security landscape (e.g., joint initiatives to combat transnational organized crimes and counter-piracy efforts).

3.4. International Cooperation

International cooperation on governance of AI in security and defence would be crucial to addressing international security challenges, and ensure the responsible development, deployment and use of these technologies. A state implementing a national strategy should take stock of on-going governance processes and discussions and initiatives at the international level, both within the United Nations and outside. Building on this assessment, the state should consider its role in the international community, and engage accordingly through active participation in international forums, supporting international institutions, and leading on state-led initiatives aimed at fostering global cooperation on AI in security and defence.[8] Transparency will form a key part in underpinning international cooperation; information-sharing will, in this sense, constitute a key confidence-building measure to signal the state's commitment to the safe and responsible development of AI technologies to other players.[9]

4. Lessons from and Convergence with Select Fields

In the development, adoption, implementation and review of a national strategy on AI in security and defence, states ought to consider lessons and good practices from other security and technological fields. These comparative analyses must, however, be done with a thorough understanding of the unique characteristics of AI and, thus, its inherent differences with the other security and technological fields. Areas of technological and governance convergence between security fields should also be considered and addressed as appropriate.

4.1. The Cybersecurity Field

A state implementing a national strategy may consider the extent to which successful governance models in the cybersecurity field can be applied to AI in security and defence. These include export control models for intangible systems, the establishment of emergency response teams, and confidence-building measures. In addition, the state should consider areas of convergence between the AI and cyber fields, and should address possible risks and issues that may arise in its national strategy. Opportunities that may arise from this convergence should also be considered.

4.2. The Nuclear Security Field

A state implementing a national strategy may consider the extent to which lessons may be

---

[8] Examples of such processes include the REAIM Summit and adjacent processes, initially led by the Netherlands and the Republic of Korea and more recently also championed by Singapore, Kenya and the United Kingdom; as well as the United States' Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, which has been endorsed by a number of states and for which a series of cross-regional thematic working groups has been established.

[9] Ioana Puscas, *Confidence-Building Measures for Artificial Intelligence: A Multilateral Perspective* (Geneva: UNIDIR, 2024), https://unidir.org/wp-content/uploads/2024/07/UNIDIR-Confidence_Building_Measures_Artificial_Intelligence-Multilateral_Perspective.pdf.

drawn from nuclear governance, cognizant of the inherent technical differences between AI and nuclear technologies. These include safety frameworks, as well as approaches to delineating peaceful from non-peaceful uses. In addition, the state should consider areas of convergence between the AI and nuclear security fields, and should address possible risks and issues that may arise in its national strategy. Opportunities that may arise from this convergence should also be considered.

4.3. The Biological and Chemical Security Field

A state implementing a national strategy may consider the extent to which lessons may be drawn from the biological and chemical security field. Their general-purpose character of both AI and biological and chemical technologies, along with the risks stemming from their inherently dual-use natures, provide much room for cross-pollination and mutual learning for key issues including controlled access and non-proliferation. In addition, the state should consider areas of convergence between the AI and the biological and chemical security fields, and should address possible risks and issues that may arise in its national strategy. Opportunities that may arise from this convergence should also be considered.

4.4. Civilian AI Governance

As governance frameworks surrounding civilian AI applications emerge and increasingly proliferate, a state implementing a national strategy should think of opportunities for cross-pollination between the civilian domain and that of security and defence. Among areas of overlap, data governance and dual-use technologies are particularly important to explore and address in a national strategy document, considering both opportunities and risks in addition to offering a possible framework to delineate between civilian and non-civilian applications.

CHAPTER II.

Substantive Guidelines

This second series of guidelines for the development, adoption, implementation and review of national strategies on AI in security and defence is substantive. Echoing the approach to the first, procedural guidelines, these substantive guidelines do not seek to be prescriptive, but rather provide states and key stakeholders with an overview of the main considerations that, albeit non-exhaustive, would benefit from extensive reflections and, as appropriate and necessary, solutions to support the development, adoption, implementation and review of national strategies.

1. Data Practices

Data constitutes the lifeblood of AI technologies. In the light of the sensitivities of security and defence applications and the complex regulatory implications that surround them, states should reflect on governance approaches to data practices in their national strategies on AI in security and defence.[10]

1.1. Data-Collection Practices

A state developing a national strategy should ensure that data-collection practices for development, training and testing of AI are lawful. All data should be collected, processed and stored in compliance with the applicable data-protection laws, cognizant of the relevant nuances of security and defence applications. The state should clarify how existing data-protection regimes apply in such contexts and, if necessary, develop its interpretation of the law for specific use-cases in security and defence. The state should, in parallel, establish oversight and accountability mechanisms to ensure that data collection is done within the confines of the law. As such, data-governance frameworks with clear protocols for data collection in specific security and defence use-cases should be developed, with compliance as a guiding principle framing the state's efforts in this space.

1.2. Data Hygiene

As part of its national strategy, a state should develop and implement guidelines surrounding data hygiene practices, thus ensuring the reliability of AI systems with regards to operational effectiveness, legal compliance and ethics. Identifying and implementing good practices to ensure the hygiene of training and testing data sets would be particularly important to ensure the responsible development, deployment and use of AI in security and defence. These guidelines should also include requirements for cleaning, validation and updating of data sets on a regular basis, ensuring their veracity and, thus, the accuracy of AI systems.

1.3. Appropriateness of Training Data

The quality of training data is critical for the performance, accuracy and reliability of an AI system. A state implementing a national strategy should ensure that the AI systems that it is developing, acquiring, testing, adopting or using for security and defence applications are trained on data sets of appropriate quality, ensuring their diversity, veracity and representation. As such, the state should develop evaluation benchmarks and metrics to measure the appropriateness of training data, to be developed with multidisciplinary input from the relevant stakeholders. The state should also consider the establishment of a national repository of approved training data sets that have been evaluated against legal and ethical requirements and relevant standards, thus ensuring traceability and consistency in the quality of training data used across various applications.

---

[10] Yasmin Afina and Sarah Grand-Clément, "Bytes and Battles: Inclusion of Data Governance in Responsible Military AI Discussions", Centre for International Governance Innovation (CIGI), forthcoming 2024.

1.4. Appropriateness of Testing Data

The quality of testing data is also critical for the performance, accuracy and reliability of an AI system. Testing data should be different and remain separate from training data. A state implementing a national strategy should develop evaluation benchmarks and metrics to measure the appropriateness of testing data, ensuring that it accurately reflects, to the extent possible, the operational environment in which the AI system will be deployed. This will, in turn, enable the conduct of comprehensive, robust and reliable testing and evaluation of AI systems in security and defence. The state should also deliberate on the question of the production of, and subsequent access to, testing data. In particular, access to testing data by technology developers should be framed and considered carefully, especially to avoid the optimization of system performance solely against testing data at the expense of downstream deployment and reliability for real-world use.

1.5. Proxy Data Use

Proxy data, or proxy indicators, are often used when the actual data (i.e., direct indicators) required for training an AI model is either unavailable or too sensitive to use. Use of proxy data will be particularly important and prominent in security and defence applications, where direct indicators may not always be available. For example, ISR efforts rely heavily on proxy data. A state developing a national strategy should consider the opportunities and, conversely, risks stemming from proxy data use for the training, testing and deployment of AI in security and defence, in close consultation with the relevant communities across disciplines. The state should then consider the establishment of guidelines or, at least, good practices that frame the responsible use of proxy data in the context of AI in security and defence. Oversight mechanisms, verification measures and accountability frameworks for the use of proxy data should also be considered and developed by the state as appropriate.

1.6. Synthetic Data Use

The availability, quality and diversity of data have long been raised as key challenges to the development, deployment and use of AI in security and defence.[11] While synthetic data could, in principle, help overcome the limitations stemming from insufficient and unavailable data sets for the training, testing and deployment of AI systems, a host of concerns arise with regards to their quality and representativeness. As such, each state should develop and adopt guidelines or, at least, good practices for the responsible use of synthetic data in the context of AI in security and defence. Oversight mechanisms, verification measures and accountability frameworks over the use of synthetic data should also be considered and developed by the state as appropriate. In parallel, the state should consider the specific use-cases where synthetic data use would offer a cutting-edge advantage in security and defence and, at the same time, address the root causes of insufficient data for the training, testing and deployment of these systems.

1.7. Data Governance

A state implementing a national strategy should put in place a framework with clear guidelines on the governance of data in the context of AI applications in security and defence. To this end, the state should review legal frameworks, policies and other regulatory instruments in place surrounding data governance and the

[11] Harry Deng, *Exploring Synthetic Data for Artificial Intelligence and Autonomous Systems: A Primer* (Geneva: UNIDIR, 2023), https://unidir.org/publication/exploring-synthetic-data-for-artificial-intelligence-and-autonomous-systems-a-primer.

extent to which they are relevant and appropriate for AI development, deployment and use in security and defence. Notably, the state should consider data privacy issues as well as existing practices, processes and mechanisms in place for storing and sharing data, including in the context of cross-border data transfers with integrated autonomy or automation.

2. Implications of Machine Learning Use

The development, deployment and use of machine learning for security and defence applications bring to the fore a series of considerations specific to the nature of these technologies. National strategies on AI in security and defence should thus account for the unique characteristics and considerations of these technologies.

2.1. Black Box

The "black box" nature of AI systems – where decision-making processes are not transparent and the calculations leading to a system's outputs remain opaque – is a significant concern shared by a host of stakeholders in the light of these technologies' development, deployment and use in security and defence. A state developing a national strategy should ensure that AI systems are designed, to the extent possible, with explainability features to ensure and preserve traceability and accountability, especially for applications involving high risks and high consequences (e.g., target identification in military settings). At the very least, the state should consider the management and oversight of applications where the black box issue is inevitable, and should clarify the accountability and responsibility of the relevant actors in the development, deployment and use of these technologies.

2.2. Probabilistic Nature of AI Technologies and Accuracy Rates

AI technologies have an inherently probabilistic nature: their outcomes are based on likelihood and probabilities over certainty. Each state should thus account for the inherent uncertainties brought by AI use in security and defence, particularly with regards to risk assessments ahead of deployment. As such, benchmarks and standards to evaluate a system's accuracy and reliability against operational needs, legal compliance, and alignment with ethical guidelines will be key in addressing the probabilistic nature of AI in security and defence.

2.3. Safe Testing Environments and Practices

The robust testing of an AI system in realistic environments prior to its deployment it critical to measure and ensure its reliability. Testing must simulate the full range of operational contexts, including edge-cases that may reveal system weaknesses. The environments in which these tests are conducted, and adjacent practices, must be controlled and safe. To this end, a state implementing a national strategy should thus invest in the financial, human and technological resources required for safe testing environments and practices. These should include regulatory sandboxes specifically dedicated to security and defence applications, providing a controlled environment for the testing and evaluation of new and emerging technologies. Good practices for the documentation and

oversight of testing practices should also be considered and prioritized.

2.4. Safe Evaluation Environments and Practices

Good evaluation practices should go beyond initial testing, extending throughout the life cycle of an AI system. A state implementing a national strategy should conduct a periodic assessment of the performance, legality and compliance of AI technologies, and the ethical implications of these systems in operational use. The state should develop a framework for continuous evaluation of AI systems and should invest in the technological solutions required to conduct such monitoring (e.g., through audit frameworks and systems).

2.5. Transparency in Testing and Evaluation Processes

Transparency in the testing and evaluation process of AI systems builds trust and confidence, both within a state and among regional and international partners. Sharing non-sensitive aspects of AI testing and evaluation processes would constitute a key component of information exchange as a confidence-building measure at both the domestic and international levels.[12] A state implementing a national strategy should thus consider the creation and publication of transparency reports on the testing and evaluation of AI systems in security and defence, while ensuring that sensitive and classified information is preserved to ensure the integrity of the state's national security.

2.6. Open-Source AI

The application of open-source practices in AI could, in principle, foster innovation. An increasing number of security and defence agencies have already adopted, or at least considered the adoption of, open-source AI. Yet, these practices could also present severe risks, particularly in security and defence in the light of the sensitivities of such applications. Hence, a state implementing a national strategy should develop clear policies and guidelines that balance the need for open innovation, leveraging the opportunities that open-source practices offer, while ensuring that the sensitivities that stem from security and defence applications are considered and addressed accordingly.

---

[12] Puscas, *Confidence-Building Measures for Artificial Intelligence*.

3. Public–Private Partnerships

Public–private partnerships are key to fostering responsible innovation in AI, including in the security and defence sector. States should thus explore ways in which such partnerships can be leveraged to advance the implementation of national strategies, while ensuring such collaboration is effective, meaningful and sustainable.

3.1. Distribution of Roles and Responsibilities

In the implementation of its national strategy, a state should establish frameworks that provide a clear definition and distribution of roles and responsibilities between the public and private sector. While the enforcement and oversight of the national strategy should remain the sovereign prerogative of the state, it should also consider the role that the private sector could play in implementation. The state should provide for the accountability and responsibility of the private sector in the context of the development, deployment and use of AI in security and defence; the development, adoption, implementation and review of the national strategy must thus include these considerations. In addition, the state should also consider the formulation of performance measurement and evaluation, with specific guidelines to evaluate and measure the success of public–private partnerships.

3.2. Foster Research and Development Ecosystems

A thriving research and development ecosystem is critical for responsible AI innovation. A state implementing a national strategy should promote collaboration between government, academia and industry as appropriate in the context of the development, deployment and use of AI in security and defence. Through the establishment of national AI innovation frameworks and structures, the state should incentivize research and development, in addition to mobilizing and allocating the financial, human and technological resources required to foster its national research and development ecosystem.

3.3. Intellectual Property

A state implementing a national strategy should consider the implications that public–private partnerships may have for intellectual property rights, particularly with regards to the development of dual-use technologies. The state should provide guidelines to ensure clear and defined ownership and licensing rights for the intellectual property developed within public–private partnerships, and the extent to which intellectual property rights may be leveraged to protect sensitive information and technologies.

4. Technological Transfer

States should outline, in their national strategies, their approaches to technological transfer, whether in the context of procurement or sales of AI in security and defence.

4.1. Technological Transfer, Equity and Proliferation Risks

Technological transfer, especially with regards to AI in security and defence, may be driven by a number of considerations, including procurement, sales, assistance and joint efforts with allied states. Yet, when implementing a national strategy, a state should consider how to approach this issue while considering risks surrounding technological proliferation and potential misuse. Considerations should extend beyond the transfer of the technology itself, and cover "package deals" that include training, as well as the technical and broader assistance that can be offered to ensure that the technology is successfully and responsibly adopted and integrated into the recipient state's military capabilities. The state should thus develop a framework to ensure that it is able to maintain control over the transfer of these sensitive technologies; however, these concerns must also be counterbalanced by the need to promote equitable access to AI capabilities for responsible actors. The state should thus establish comprehensive policies to frame the technological transfer of AI systems in security and defence to foreign governments, private entities or international partners.

4.2. Procurement from Foreign Governments

A state implementing a national strategy should carefully evaluate its approach to the procurement of AI capabilities from foreign governments. The state should thus develop a framework to govern such procurements, ensuring that they meet security and operational needs in addition to legal compliance and alignment with ethical guidelines. The state should ensure that clear and robust processes are in place for local acceptance tests along with pre- and post-sale feedback mechanisms with the foreign seller, and it should invest in the solutions needed to ensure the safe and secure interoperability of these new technologies with existing systems.

4.3. Sales to Foreign Governments

In implementing a national strategy, a state should carefully evaluate its approach to sales of AI capabilities for security and defence applications to foreign governments. The state should thus develop a framework to govern such sales, ensuring that they are conducted in compliance with existing international arms control obligations and in alignment with national security interests. Clear and robust processes must be in place to ensure transparency in these transactions and ensure a degree of oversight to prevent misuse in ways that would undermine regional security or, more generally, international peace and security. Such a framework should dedicate specific considerations to both government-to-government sales and business-to-government sales (i.e., when a private actor operating on a state's territory sells AI-enabled capabilities to foreign governments).

4.4. Sales to Non-Governmental Entities

Sales of AI applications for security and defence to non-government entities should be strictly regulated through the establishment

of a framework governing such transactions. In order to reduce risks of misuse, each state should establish and implement thorough controls and oversight mechanisms to ensure that these technologies are neither used nor repurposed to enable the violation of international law and in such ways that would compromise national security. The state should thoroughly vet any non-governmental entity seeking to purchase AI applications for security and defence, evaluating their intended use, the entity's capacity for responsible management and use, and its adherence to applicable laws. Each state should thus consider the establishment of a licensing body to examine, review and deliberate on (i.e., either approve or reject) sales of such capabilities to non-governmental entities. Specific contractual clauses should also be included to ensure and maintain accountability and responsibility pre- and post-sales.

4.5. Oversight and Verification

In the light of the sensitive nature of AI technologies, particularly those with applications in security and defence, each state should consider the establishment of oversight and verification mechanisms in the context of technological transfers – whether the state acts as the supplier or purchaser of these capabilities. The state should consider the establishment of an independent oversight body mandated to review AI technology transfers. Its establishment, composition and governance should be cognizant of the sensitivities surrounding this topic, while ensuring such a body has the capacity and power to conduct investigations and report on any transfers that may pose a risk to national, regional or international security.

5. Life Cycle Management

A national strategy on AI in security and defence should outline the ways in which the state intends to conduct the management of these technologies' life cycle. While there is no universally accepted model of an AI technology's life cycle and its stages, management of the life cycle provides states with clear entry points to promote responsible behaviour and practices from the development to the decommissioning of AI technologies in security and defence.[13]

5.1. Identification of Relevant Life Cycle Stages

Effective life cycle management would require the clear identification of all stages in an AI technology's life cycle, from design and development to decommissioning. A state implementing a national strategy should consider that the life cycle of most, if not all, AI technologies is neither linear nor necessarily cyclic in a straightforward manner. As such, the state should design policies and frameworks that identify and outline each life cycle

[13] Yasmin Afina and Giacomo Persi Paoli, *Governance of Artificial Intelligence in the Military Domain: A Multi-Stakeholder Perspective on Priority Areas* (GENEVA: UNIDIR, 2024), https://unidir.org/publication/governance-of-artificial-intelligence-in-the-military-domain-a-multi-stakeholder-perspective-on-priority-areas/.

stage, and incorporate good practices from both military and civilian use of AI to promote responsible behaviour and practices. State agencies may have their own definitions, interpretations and approaches to the life cycle of these technologies, depending on their respective mandate, jurisdiction and policies; however, a fragmented approach to life cycle management across agencies may carry risks with regards to vulnerabilities, oversight (or lack thereof) and interoperability. As such, the state should develop and implement a national life cycle management framework for AI in security and defence to ensure alignment at the national level.

5.2. Identification of Relevant Actors Across Life Cycle Stages

The complexity of AI systems requires the involvement and intervention of various actors, both public and private, across the technology's life cycle. Each actor – from developers to testing and evaluation professionals, lawyers, and end-users – should have clearly defined and distributed roles and responsibilities at each stage of the technology's life cycle. A state implementing a national strategy should thus ensure the involvement of all relevant stakeholders, as appropriate, in the development, adoption, implementation and review of the national strategy to ensure that no actor is overlooked across the technology's life cycle. Clear communication channels and coordination mechanisms should be established among all stakeholders, including the facilitation of regular meetings that transcend state agencies and disciplines and, as appropriate, incorporate the viewpoints of both governmental and non-governmental entities, including the private sector.

5.3. Distribution of Roles and Responsibilities Across the Technology's Life Cycle

A clear definition, delineation and distribution of roles and responsibilities across the technology's life cycle would prevent gaps in oversight and ensure that AI systems are managed effectively. A state implementing a national strategy should thus adopt and establish a clear organizational structure, with the appropriate resources allocated, to ensure effective coordination between actors, as well as the implementation and oversight of these distributed roles and responsibilities.

5.4. Considerations for the Technology's Procurement

The procurement of AI systems should be managed and framed carefully to ensure alignment with national, regional and international policies, applicable laws, as well as ethical guidelines. A state implementing a national strategy should thus take stock of current procurement policies and frameworks, review procurement practices for AI capabilities, and evaluate their suitability to ensure their security, reliability and compliance. The state should establish procurement guidelines that include testing and evaluation benchmarks that combine considerations of compliance with applicable laws, ethical guidelines and operational needs. Local acceptance test benchmarks will be particularly critical for capabilities purchased externally (i.e., from other states or from foreign industries) and off-the-shelves capabilities due to their untailored development and training. A periodic review of these procurement guidelines should be conducted to factor in technological progress and the emergence of new norms and principles.

5.5. Considerations for the Technology's Development

The design and development stages of AI technologies in security and defence must be centred around legal compliance and ethical guidelines. The translation of legal requirements and ethical considerations into concrete action for the design, development and training of AI systems will be critical. These could include, for example, good practices for compliance with international humanitarian law or data practices for the training and testing of AI-enabled target-identification systems. Additionally, each state should ensure that safety and security considerations are incorporated from the early stages of the development of these technologies, including identification and mitigation of bias and resilience against adversarial attacks.

5.6. Considerations for the Technology's Testing and Evaluation

The testing and evaluation of AI technologies in security and defence will play a key role in ensuring reliability and the alignment of these systems with a state's policies, applicable laws and ethical guidelines. From an operational standpoint, testing and evaluation are critical to ensure that the systems perform as intended and as expected under the appropriate conditions. Beyond performance, testing and evaluation are also critical in verifying the implementation and effectiveness of key principles (e.g., human control, judgement, oversight and involvement, legal compliance, and alignment with ethical principles). The state should thus not only develop comprehensive testing and evaluation frameworks with clear protocols as part of its national strategy on AI in security and defence, it must also invest in and allocate the resources necessary to facilitate thorough testing and evaluation (e.g., through simulations, adversarial stress-testing for edge cases, and the identification of and planning for potential failure scenarios).

5.7. Considerations for the Technology's Adoption and Deployment

The adoption and deployment of an AI technology in security and defence should be done in a phased and controlled manner, to ensure that it is integrated safely and securely into existing systems. The effective integration and interoperability should be tested, monitored and evaluated periodically to ensure consistent reliability and predictability. Each state should develop and adopt clear procedures for the introduction of AI into operational settings and address issues that may arise post-deployment.

5.8. Considerations for the Technology's Use

A national strategy should consider the relevant operational, policy, legal and ethical considerations and implications of the use of AI in security and defence. These include legal compliance, accountability and the human element, specifically the ways in which the state should ensure, when appropriate, human control, judgment, oversight or involvement. A risks-based approach to framing and governing the use of AI technologies in security and defence should be considered, as well as the appropriate measures necessary to mitigate risks stemming from their operational use.

5.9. Considerations for the Technology's End of Life

The effective management of an AI system's end of life is critical for the responsible development, deployment and use of AI in security and defence. A national strategy should outline clear procedures for the retirement and decommissioning of AI systems, including secure data destruction and the safe and secure recycling or repurposing of components. Additionally, the state should consider the management of second-hand sales and technological transfer to other states, and should address concerns and risks surrounding non-proliferation, reliability and accountability of re-purposed technologies that have been previously used and

subsequently decommissioned. To this end, a designated agency or body should be identified and designated as responsible for overseeing the safe and secure management of a technology's end of life.

6. Human Resources

While the development, adoption, implementation and review of national strategies on AI in security and defence will require substantial financial resources, states should not overlook the importance of considering and investing in the human resources required to ensure these strategies' effectiveness.

6.1. Training and Retention of AI Talent

Each state should cultivate and retain AI talent, and address the difficulties encountered in the recruitment and retention of qualified personnel. In close collaboration with academic and research institutions and the private sector, the state should invest in and establish long-term programmes to form a solid AI talent pool, in addition to providing competitive salaries and opportunities for professional development and advancement in the sector. The state should also consider the development of national AI training programmes and curriculums linked to career pathways specifically in the security and defence sector. Through the establishment of partnerships with academic and research institutions and the private sector, these programmes would create and facilitate specialized AI courses focused on security and defence applications.

6.2. Capacity-Building and Awareness-Raising of Developers

Developers of AI technologies in security and defence should not only be well-versed in technical skills, each state should also ensure their awareness of, and literacy with regards to ethical, legal and operational standards, considerations and obligations. The state should thus facilitate capacity-building efforts to ensure that developers undergo ethical and legal training specific to the development of AI applications in security and defence. The state should also consider incentivization for such training, including through certifications.

6.3. Capacity-Building, Upskilling and Awareness-Raising of Users

A state implementing a national strategy should invest in the intended users' technical, legal and ethical capacity and knowledge. The state should mobilize the appropriate resources required to ensure that all personnel using AI systems in security and defence receive adequate training on not only the functionality and parameters of the systems, but also the technical limitations of the systems in use, as well as the relevant legal and ethical considerations stemming from the deployment and use of these systems. Ultimately, users should be able to conduct informed risk assessments and, when operating the system, should be well-informed of its characteristics, limitations, and the legal and ethical considerations relevant to specific use-cases. Users should also be aware of the institutional architecture in which they operate, including the relevant bodies, agencies and stakeholders, oversight mechanisms, and accountability frameworks in place.

6.4. Capacity-Building and Awareness-Raising of Policy and Regulatory Entities

Policymakers and regulators must be equipped with a deep and thorough understanding of AI technologies and their inherent complexities in order to ensure the development and implementation of robust, adaptive and evidence-based governance frameworks. A state developing a national strategy should undertake a comprehensive assessment of internal capacity and knowledge gaps across governmental agencies, and subsequently invest in training for the policymaking and regulatory communities to reinforce their understanding of the technical, legal and ethical dimensions of AI in security and defence. Such training should be conducted periodically, thus ensuring policy and regulatory entities are well-informed of the technological progress and subsequent risks stemming from the development, deployment and use of AI in security and defence.

6.5. Academic and Civil Society Engagement

In implementing a national strategy, a state should establish partnerships with academic institutions for research and development in AI, thus ensuring that the state remains at the cutting edge of AI innovation. Academic and civil society engagement should also be facilitated to ensure expert and multi-stakeholder input on the policy, legal and ethical implications of AI in security and defence. To this end, the state should think of incentivization, including through the funding of research by academic institutions and civil society organizations, and the establishment of frameworks to facilitate such engagement in a consistent and sustained manner.

6.6. Industry Engagement

Acknowledging the pioneering role that industries play in driving AI innovation, research and development, including in security and defence, a state implementing a national strategy should engage closely with the private sector. Public–private partnerships will be key in ensuring that the development, deployment and use of AI in security and defence are done in a responsible, safe and secure manner. The state should thus consider the creation and maintenance of formal and informal channels and platforms to facilitate such industry engagement. Beyond establishing dialogue and fostering collaboration opportunities, industry engagement would also be critical in ensuring their alignment and adherence to national policies, international law and ethical principles. To this end, incentivization will be key, which each state should reflect on and foster through its national strategy.

7. Legal Compliance

All states are bound by international law, which includes international humanitarian law, international human rights law and public international law. As states grapple with the governance of the development, deployment and use of AI in security and defence, national approaches and policies must remain within the boundaries set by the law. As such, states should set compliance at the heart of their national strategies, and outline how they interpret and approach certain dispositions of the law. States should also outline, in their national strategies, the concrete measures that they intend to adopt and implement in order to foster compliance with international law (e.g., through the conduct of iterative legal reviews).

7.1. Compliance with International Humanitarian Law

Each state must ensure that the development, deployment and use of AI in security and defence comply with international humanitarian law. The latter is applicable in international and non-international armed conflict; this, however, does not mean that IHL considerations can be omitted by states in peacetime. Good practices to foster compliance with IHL outside armed conflict (e.g., at the design and development stages of military AI capabilities, Article 36 legal reviews) should be identified and implemented. Processes should be in place to facilitate the conduct of legal reviews while procuring, testing, evaluating and adopting AI systems, with clear guidelines and benchmarks to measure compliance. Processes should also be in place for the conduct of periodic legal reviews to ensure consistent compliance with IHL. As such, each state should also invest in the development of solutions to monitor and audit a system's ability to comply throughout its life cycle and, as necessary, flag modifications in the system's parameters and performance that may affect its ability to operate within the boundaries of IHL.

7.2. Compliance with International Human Rights Law

The development, deployment and use of AI technologies in security and defence must also comply with international human rights law, both in conflict and peacetime. Each state should thus clarify the IHRL implications of the development, deployment and use of these technologies, particularly with regards to the right to life, freedom of expression and the prohibition of arbitrary detention. The state should also outline in its national strategy how it intends to foster compliance with IHRL, including through due process and legal review mechanisms. For applications in armed conflict, the state should reflect on the applicable obligations in both IHL and IHRL, and the interplay between both bodies of law in the context of AI development, deployment and use. Capacity-building should also be prioritized, notably to sensitize and raise awareness among public bodies using, or at least expected to use, AI technologies in security and defence, about their IHRL obligations and implementation mechanisms. There should also be awareness-raising with regards to human rights obligations among non-state stakeholders, notably the private sector (e.g., industries developing AI capabilities in security and defence).

7.3. Compliance with Public International Law

Public international law governs state behaviour and relations, including in the context of AI development, deployment and use in security and defence. As such, when developing a national strategy, a state should consider the public international law implications of AI, possible risks and mitigation measures, as well as its interpretation of a number of contentious aspects in this space. Such reflections are particularly important in the light of the possible integration of AI into decision-making processes surrounding the resort to and use of force, and subsequent implications on *jus ad bellum*. Considerations should include state responsibility and the question of attribution and the implications for state sovereignty of AI deployment and use (e.g., in AI-enabled trans-border ISR activities).

7.4. Compliance with Other Bodies of International Law

When developing, adopting, implementing or reviewing a national strategy on AI in security and defence, a state should also consider the applicability and subsequent application of other bodies of public international law. These include international criminal law, international maritime law, international space law and international refugee law. Building on these reflections, the state should take stock of the statutory and customary obligations that it is bound by in each of these bodies of law, and adopt the appropriate measures needed to foster compliance in the context of the development, deployment and use of AI in security and defence applications.

7.5. Compliance Oversight and Evaluation Mechanisms

Each state should clarify its position on the interpretation of international law in the context of AI development, deployment and use in security and defence. Additionally, each state should establish oversight mechanisms to ensure that these technologies remain consistently compliant with applicable laws. The conduct of periodic evaluations, and the establishment of a process to ensure documentation and reporting – whether public or internal – of compliance efforts for each branch of international law should be considered and facilitated.

8. Ethics

Ethics sit at the heart of the approach of many states to the governance of AI in security and defence. Yet, unlike international law, ethics are not binding and so views, interpretations and prioritization may diverge widely across states. In addition, there is no established mechanism for ethical accountability and responsibility. National strategies thus present an opportunity for states to outline their approaches to key ethical issues and how they translate into policy and legal considerations.

8.1. Gender Bias

Without the appropriate safeguards and risk-mitigation measures in place, AI systems could reinforce gender biases that will have

severe security and defence implications.[14] Fairness and inclusivity should be considered as an approach to address the risks stemming from gender bias. Thus, when developing a national strategy, a state should consider outlining possible approaches and solutions to addressing these risks. These could include benchmarks and guidelines on the use of diverse and representative data sets, the establishment of algorithmic audits, and corrective measures when biases are detected. Gender perspectives should be considered and integrated, as appropriate, at every stage of the life cycle of an AI technology, from design and conceptualization to operational use.

8.2. Racial Bias

Left unaddressed, racial bias in AI can have severe legal and ethical repercussions. In the context of both security applications (e.g., law enforcement) and defence (e.g., military targeting), racial biases could lead to the violation of international humanitarian law and international human rights law. Each state should thus ensure that AI systems used in these contexts are subject to thorough scrutiny and review processes to avoid the exacerbation of issues stemming from racial bias. The regular conduct of audits and reviews of the system's performance will be critical, particularly when these systems are used for operations in areas of diverse demography. The introduction of review and oversight mechanisms, along with a systematic compliance assessment, should be prioritized.

8.3. Other Forms of Discrimination

The development, deployment and use of AI in security and defence should prevent the perpetuation and exacerbation of any form of discrimination, including those based on disability, age or socioeconomic status. A state developing a national strategy should ensure that it explicitly addresses these potential biases through concrete recommendations and solutions, including through robust testing, evaluation and validation processes to prevent discriminatory outcomes downstream. The incorporation of ethics into the design of AI technologies for security and defence should be prioritized, with input from ethicists, experts and representatives of minorities and other groups most at risk of algorithmic discrimination.

8.4. Fairness

Fairness in AI decision-making is critical for maintaining public trust and governance frameworks that are centred on ethics; this is further accentuated for AI applications in security and defence. Each state should ensure that AI systems under development, in procurement and in use are not only transparent, but that considerations around fairness have been embedded into the parameters of the systems' learning, development and testing stages. A periodic review of societal impacts and fairness evaluation should be facilitated, with multi-stakeholder input as appropriate.

8.5. Explainability and Traceability

When a state reviews key ethical factors in the development, adoption, implementation and review of its national strategy on AI in security and defence, particular attention should be dedicated to explainability and traceability. In the light of the black box and inherently probabilistic nature of AI systems, human operators should be in a position where they are able to

---

[14] Katherine Chandler, *Does Military AI Have Gender? Understanding Bias and Promoting Ethical Approaches in Military Applications of AI* (Geneva: UNIDIR, 2021), https://unidir.org/publication/does-military-ai-have-gender-understanding-bias-and-promoting-ethical-approaches-in-military-applications-of-ai/.

explain the rationale behind their decision to deploy and use AI systems in specific contexts. To this end, these technologies should have, to the extent possible and appropriate, methods and processes in place for their explainability, that is, for users to comprehend the results and outputs created by the system. Traceability – that is, the ability to understand and trace back the processes behind the development of the systems (e.g., source of training and testing data, testing and evaluation metrics, the algorithm's key parameters) – will, in this sense, be important to consider. Traceability and explainability would not only be important for risk assessments undertaken by the human operator; they would also allow, as appropriate, human operators to intervene when necessary, particularly in situations where AI-driven decision-making may contravene ethical or legal requirements. As such, the state should consider the extent to which explainability and traceability could, or should, be part of procurement processes.

9. Defence Applications of AI

The integration of artificial intelligence in defence presents a host of opportunities, from back-end support to increased autonomy in weapon systems and enhanced ISR capabilities.[15] At the international level, the governance of defence applications of AI is high on states' agendas to ensure their responsible development, deployment and use, and cognizant of their associated risks, international law obligations and ethical guidelines.

9.1. Scope of the Defence Applications of AI

Defining the scope of AI applications in defence will be critical for the development, adoption, implementation and review of national strategies on AI in security and defence. A clear and defined scope would prevent a fragmented approach across governmental agencies, in addition to ensuring a clear distribution of roles and responsibilities in the defence ecosystem. Each state should consider the development, deployment and use of AI for both offensive and defensive purposes. Furthermore, both combat and non-combat functions should be considered, in the light of their inherently different policy, legal and ethical implications. Each state should also consider the boundaries between the "defence" and "security" realms, and which regulatory instruments and bodies have jurisdiction over these applications.

9.2. Integration into Weapon Systems

A state developing a national strategy should address both the opportunities and the risks that stem from integration of AI into weapon systems. The strategy presents an opportunity for that state to outline the key guiding principles framing its approaches and priorities in the governance of the integration of AI into weapon systems. Rationales for permissible use-cases, limitations and eventual prohibitions should be included in the national strategy and substantiated, notably on the

[15] Afina, *The Global Kaleidoscope of Military AI Governance.*

basis of international law and ethical guidelines. Operational guidelines should also be developed with regards to the development, deployment and use of AI-enabled weapon systems, including specific rules of engagement when these systems are in use. Frameworks surrounding their pre-deployment stages, grounded in compliance with international law and ethical guidelines, should also be developed, including with regards to procurement practices and testing and evaluation. The state should also consider timely audits and the establishment of monitoring systems, in addition to recurrent legal reviews to ensure these systems' consistent compliance with international law, particularly international humanitarian law and international human rights law.

9.3. Integration into Non-Weapon Systems

AI integration into defence applications extends beyond weapon systems, including for enhanced ISR capabilities, training and simulation, and logistics support.[16] While these technologies offer significant operational advantages, these systems must be developed, deployed and used within the limits imposed by international law and ethical guidelines. A state developing a national strategy should take stock of existing defence applications of AI outside weapon systems, and subsequently develop a clear governance framework surrounding their development, deployment and use. The establishment of training benchmarks, rigorous testing and evaluation metrics, and good practices for their development and use will be key to the successful implementation of national strategies. The state should also consider timely audits and the establishment of a monitoring system, in addition to recurrent legal reviews to ensure that the consistent compliance of these weapon systems with international law, particularly IHL and IHRL.

9.4. Legal Compliance

Ensuring legal compliance in the context of AI development, deployment and use in defence will be critical and should lie at the heart of a national strategy. When developing a strategy, a state should clarify its interpretation of international law in the context of AI in defence, and develop a national position that captures and crystallizes its approach.[17] In addition, the state should adopt a comprehensive legal review mechanism to ensure the consistent compliance of AI development, deployment and use with international law, particularly IHL and IHRL. Such reviews must be conducted periodically throughout the technology's life cycle.

9.5. Key Arms Control and Disarmament Considerations

A national strategy should consider the implications of AI applications in defence for key arms control and disarmament frameworks, and outline the state's approach, position and ambitions in this space. These considerations may include the opportunities that AI presents to reinforce existing arms control and disarmament treaties, the associated risks, as well as the relationship and interplay between relevant international, regional and national frameworks.

---

[16] Sarah Grand-Clément, *Artificial Intelligence Beyond Weapons: Application and Impact of AI in the Military Domain* (Geneva: UNIDIR, 2023), https://unidir.org/publication/artificial-intelligence-beyond-weapons-application-and-impact-of-ai-in-the-military-domain/.

[17] For a compendium of good practices surrounding such processes in cyber, where analogies can be drawn, see UNIDIR Security and Technology Programme, *A Compendium of Good Practices: Developing a National Position on the Interpretation of International Law and State Use of ICT* (Geneva: UNIDIR, 2024), https://unidir.org/publication/a-compendium-of-good-practices-developing-a-national-position-on-the-interpretation-of-international-law-and-state-use-of-ict/.

10. Security Applications of AI

The integration of artificial intelligence in security has vast potential to support states both in law enforcement (e.g., crime analysis and investigative support) and in national security, spanning from ensuring border security, supporting intelligence analysis, to enhancing the cybersecurity and resilience of critical national infrastructure (e.g., threat detection and prevention). In certain regions, the development, integration and adoption of wider security applications of AI technologies constitute the priority over those in the defence realm.

10.1. Scope of the Security Applications of AI

For many years, states have already been unpacking the opportunities offered by artificial intelligence in security, with a number of applications in deployment and use. A state developing a national strategy should thus take a comprehensive overview of intended uses, associated risks and safeguards measures as part of the strategy. The state should also consider the extent to which "grey zone" applications would fall within the scope of security, that is, applications lying at the intersection of security and defence (e.g., deployment of military-grade ISR capabilities to support law enforcement in times of national crisis). Thresholds of acceptability, built on international, regional and domestic legal frameworks, along with ethical guidelines, should be included to ensure the responsible development, deployment and use of AI across all security applications. Regular review mechanisms should be integrated into the strategy to adapt to new and emerging security threats, risks and evolving technologies.

10.2. AI Integration into Surveillance Operations

A state that develops, adopts, integrates or uses AI in surveillance should consider applicable international legal frameworks, particularly international human rights law, and ethical guidelines. As such, oversight frameworks and mechanisms should be in place to ensure AI-enabled surveillance practices remain in compliance with IHRL (e.g., proportionality and necessity). Good practices include the establishment of an oversight body, the publication of transparency reports on AI use, and the development of accountability frameworks. Audit solutions and mechanisms should also be developed, established and subsequently implemented for the periodic review of surveillance data and practices. The state should prioritize investment in technological solutions (e.g., privacy enhancing technologies) and capacity-building (e.g., training of law enforcement and intelligence officers on responsible AI use and applicable legal frameworks).

10.3. Cross-Border Data-Sharing Practices and AI Implications

Increased use of AI in security raises novel issues and exacerbates existing sensitivities raised by cross-border data-sharing practices, particularly with regards to national sovereignty and privacy. Each state should thus consider the establishment of a national framework and bilateral agreements defining the viewed implications, parameters, conditions and limitations of AI integration into cross-border data-sharing practices. Provisions on applicable laws, safeguards and oversight should be included. Furthermore, the state should conduct a survey of possible risks stemming from AI development,

deployment and use in the context of cross-border data-sharing. Risks of unauthorized access, data interception, misuse and compromised integrity of data should, among others, be considered and addressed in such frameworks and agreements.

10.4. AI Integration into Law Enforcement Mechanisms, Practices and Operations

The development, integration, adoption and use of AI in law enforcement mechanisms, practices and operations offer a host of opportunities. However, their inherent risks must be addressed through robust oversight and accountability mechanisms. Each state should conduct a survey of existing applications, on-going procurement processes, and future plans for AI integration into law enforcement mechanisms and practices, along with a comprehensive assessment of risks and relevant actors. It should establish and implement good practices to ensure the responsible development, integration, adoption and use of AI in law enforcement. These practices include the identification and mitigation of harmful biases, the establishment of oversight, accountability and redress mechanisms, as well as the undertaking of regular audits and compliance reviews, notably in relation to international human rights law. Capacity-building should be prioritized, particularly for law enforcement officers, on the parameters and limitations of the AI-enabled tools in use.[18]

10.5. Oversight and Accountability

Each state should establish and implement rigorous oversight mechanisms for AI development, deployment and use in the security sector. Accountability frameworks and structures must be in place to review the development, procurement, deployment and use of AI in security contexts, and to ensure compliance with applicable regulatory frameworks at the national, regional and international levels. The establishment of oversight and accountability bodies should ensure that mechanisms are in place to enable relevant agencies and concerned stakeholders – from both the public and the private sectors – to provide input. Risk assessments should be conducted on a regular basis, and their findings must be integrated into the mandate of the established oversight and accountability bodies.

[18] For an example of a responsible AI toolkit in law enforcement, see United Nations Interregional Crime and Justice Research Institute (UNICRI), "The Toolkit for Responsible Artificial Intelligence Innovation in Law Enforcement", n.d., https://unicri.it/topics/Toolkit-Responsible-AI-for-Law-Enforcement-INTERPOL-UNICRI.

11. AI Integration into Critical National Infrastructure

The integration of artificial intelligence technologies into critical national infrastructure should be considered in the development, adoption, implementation and review of national strategies on AI in security and defence. As states are increasingly adopting and integrating, or at least considering the adoption of, AI technologies into critical national infrastructure, reflecting on their security implications and ensuring their resilience will be key.

11.1. Assessment of Past, Present and Future AI Integration into Critical National Infrastructure

In the development of a national strategy, a state should conduct an assessment of past, present and planned (or at least intended) integrations of AI into critical national infrastructure. Such an assessment should include a comprehensive review of the systems and hardware being used, information on their supply chains (e.g., supplier, origin of components, testing and evaluation records, procurement history), relevant actors, processes in place for auditing, maintenance plans, the systems in which they are integrated, envisioned end of life and plans for decommissioning. This would enable the state to formulate an effective strategy for the integration of AI technologies into critical national infrastructure, ensuring the compatibility and interoperability of these systems.

11.2. Risk Assessment

In its review of past, present and future AI integration into critical national infrastructure, a state should also conduct a comprehensive risk assessment. It should include consideration of risks from external factors (e.g., adversarial AI attacks), risks introduced from the integration of AI into critical national infrastructure, relevant actors, as well as risks pertaining to the supply chain of AI technologies and subsequent implications for the resilience and integrity of critical national infrastructure. These risk assessments should be conducted on a regular basis to ensure that the state remains up to date in terms of the introduction of novel risks, including those stemming from subsequent integration of AI into critical national infrastructure.

11.3. Sectoral Risk-Mitigation Measures

Upon the completion of the risk assessment, a state should develop risk-mitigation measures for each sector that may be of relevance to security and defence. These risk-mitigation measures must establish a clear distribution of roles and responsibilities for the different actors named across governmental agencies. The agency leading on risk-mitigation efforts for each sector should develop a sector-specific plan with concrete policy recommendations, pathways for implementation and oversight mechanisms. A review of such plans must be conducted on a regular basis and in consultation with relevant stakeholders, from both the public and the private sectors, to ensure the continued timeliness of the measures in place.

12. Resilience and Preparedness

In the development, adoption, implementation and review of national strategies on AI in security and defence, states should consider, and ensure, resilience and preparedness in emergency situations that stem from the use of these technologies.

12.1. Plan for Emergency Response

Each state should consider and identify what possible incidents, situations of shock, and crises may arise from the use of AI in security and defence. The assessment must be holistic in order to factor in the different facets of such incidents (e.g., scale, actors involved, harms caused, primary and secondary effects, potential geopolitical repercussions, etc.). This exercise must then lead to the careful planning of the appropriate emergency response at all levels.

12.2. Establishment of an Emergency Response Team

Upon the establishment of an emergency response plan, a state should also establish an inter-agency emergency response team, with a clear mandate and set responsibilities in case of incidents. Akin to computer emergency response teams (CERTs), the mandate of the AI emergency response team will be commensurate with the incident at hand (in terms of, e.g., scale, actors involved, harms caused, primary and secondary effects, potential geopolitical repercussions).

12.3. Distribution of Roles and Responsibilities

A national strategy should include a clear distribution of roles and responsibilities in case of incidents. This distribution should be done carefully, considering each governmental body and agency's respective authority, mandate, and available human and financial resources.

12.4. Capacity-Building

A state should prioritize capacity-building efforts to ensure that all actors involved in ensuring national resilience and preparedness have all the knowledge, guidance and resources to implement their respective roles and responsibilities. Ensuring the capacity of actors is critical both in response to incidents and for ensuring the day-to-day resilience of the state, especially if the state is considering, or already adopting, artificial intelligence technologies with security and defence implications, including in the military domain, in critical national infrastructure, and for national security purposes (e.g., law enforcement and border security).

12.5. Risk Assessment and Risk Mitigation

The development of an emergency response plan should include the conduct of a comprehensive risk assessment and the identification of mitigation measures. The latter should include the roles and responsibilities of both public and private actors, key intervention points, as well as the resources required and to be mobilized by each actor. Risk assessments and risk-mitigation plans should be conducted and reviewed on a regular basis to ensure that the risk assessments are up-to-date, and they should include the perspectives, needs and realities of all relevant stakeholders from the public and private sectors.

12.6. Incident Recording and Analysis

Each state should establish a system to record incidents, including key facts, actors, reactions and processes in place, and responses. Such a system would enable the relevant authorities to identify lessons and foster preparedness and resilience against future incidents by, for example, adapting existing protocols, doctrines and standard operating procedures. The state should allocate the resources required to ensure the processes, capacity and infrastructure needed are in place for implementation.

12.7. The Conduct of Exercises and Stress-Testing

Each state should consider the systematic conduct of exercises on a regular basis to test the existing measures, protocols, doctrines and standard operating procedures in place for incident response. These exercises would enable the state to stress-test its continued effectiveness and adapt accordingly, should the need arise. This would enable the timeliness of the processes in place, as well as consistent preparedness against new and emerging risks. Such exercises include, but should not be limited to, red-teaming and adversarial testing. The state should consider whether such exercises should be facilitated by internal or external parties, considering the value of, on the one hand, independent reviews and, on the other, sensitivities.

Conclusion and Next Steps in the Project

Although the international community is still in the early stages of developing, adopting and implementing national strategies on AI in security and defence, there is nevertheless widespread acknowledgment of the importance of these strategies for the governance of the development, deployment and use of these technologies. To this end, states must adopt a holistic approach to the development, adoption, implementation and review of these national strategies.

The above draft guidelines for the development, adoption, implementation and review of national strategies on AI in security and defence are offered by the UNIDIR project in support of states' efforts in this space. The guidelines, both procedural and substantive, do not seek to be prescriptive, but rather to provide states with a series of considerations, approaches and tools to support their respective national processes.

As UNIDIR proceeds with the consolidation of the draft guidelines and the development of commentaries on each guideline, it seeks feedback from the international community to make these resources as effective and as inclusive as possible. By the end of 2024, UNIDIR will consolidate and finalize the draft guidelines, which be integrated into its AI Policy Portal.[19] In 2025, UNIDIR will launch the second phase of this project, which will consist of developing substantive commentaries for each guideline, dissecting in further detail their importance, as well as existing approaches and thinking in both policymaking and academic scholarship. Ultimately, it is hoped that states will find these guidelines useful in supporting the development, adoption, implementation and review of their national strategies on AI in security and defence by giving access, in one place, to the main considerations and good practices that may be of relevance to their efforts.

[19] See UNIDIR Artificial Intelligence Policy Portal (AIPP), https://aipolicyportal.org/.

@unidir
/unidir
/un_disarmresearch
/unidirgeneva
/unidir

Palais des Nations
1211 Geneva, Switzerland

WWW.UNIDIR.ORG