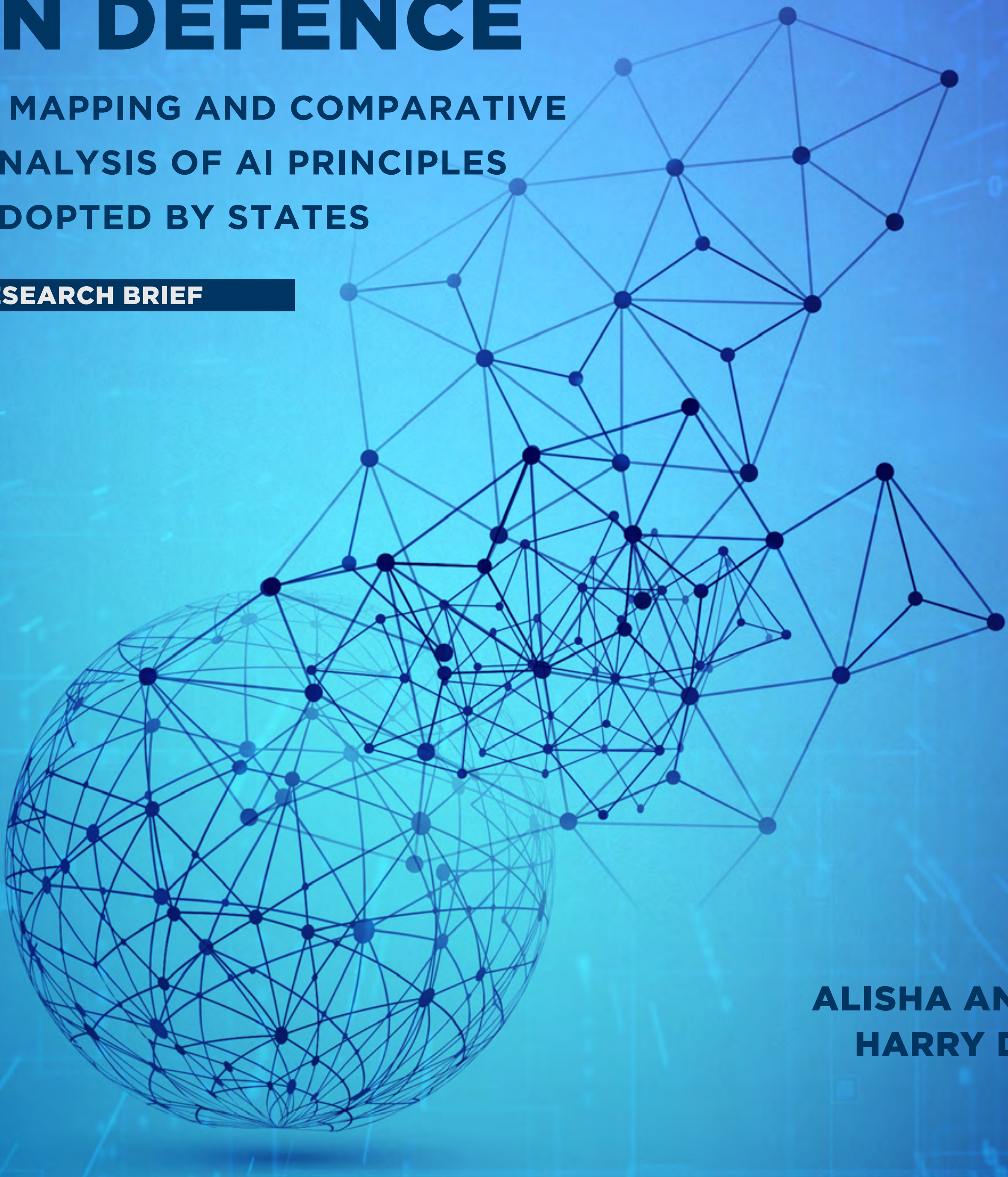
UNIDIR
UNITED NATIONS INSTITUTE
FOR DISARMAMENT RESEARCH

TOWARDS RESPONSIBLE AI IN DEFENCE

A MAPPING AND COMPARATIVE ANALYSIS OF AI PRINCIPLES ADOPTED BY STATES

RESEARCH BRIEF



ALISHA ANAND
HARRY DENG

ACKNOWLEDGEMENTS

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities. This paper was produced by the Security and Technology Programme, which is funded by the governments of Czechia, Germany, Italy, the Netherlands and Switzerland and by Microsoft. The authors wish to thank Dr Merel Ekelhof and Dr Giacomo Persi Paoli for their substantive input and guidance on this publication and Ms Belen Bianco for her research support in the initial stages of this research.

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.

CITATION

A. Anand and H. Deng, *Towards Responsible AI in Defence: A Mapping and Comparative Analysis of AI Principles Adopted by States*, Geneva, Switzerland: UNIDIR, 2023.

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.

TABLE OF CONTENTS

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 (SecTec) 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.

ABOUT THE AUTHORS

Alisha Anand is an Associate Researcher in the Security and Technology Programme at UNIDIR. Her work is focused on the international security implications of new and emerging technologies and technology governance, particularly in the field of artificial intelligence. Before joining UNIDIR, she worked on non-proliferation and export controls with the Disarmament and International Security Affairs Division of the Indian Ministry of External Affairs, the Manohar Parrikar Institute for Defence Studies and Analyses and the Federation of Indian Chambers of Commerce & Industry. She holds a master's degree in law and diplomacy from the Fletcher School, Tufts University, where she specialised in international security and international law. Follow Alisha on Twitter @AlishaAnand912.

Harry Deng is a Graduate Professional in the Security and Technology Programme at UNIDIR, where his work focuses on the international security implications of new and emerging technologies. Prior to joining UNIDIR, he worked on digital trade policy with the Centre for International Governance Innovation and on data ethics at UN-Habitat and was a cybersecurity analyst in the private sector. He holds a master's degree in global governance from the University of Waterloo, where he specialised in international political economy. Follow Harry on Twitter @hwrdeng.

ABBREVIATIONS AND ACRONYMS

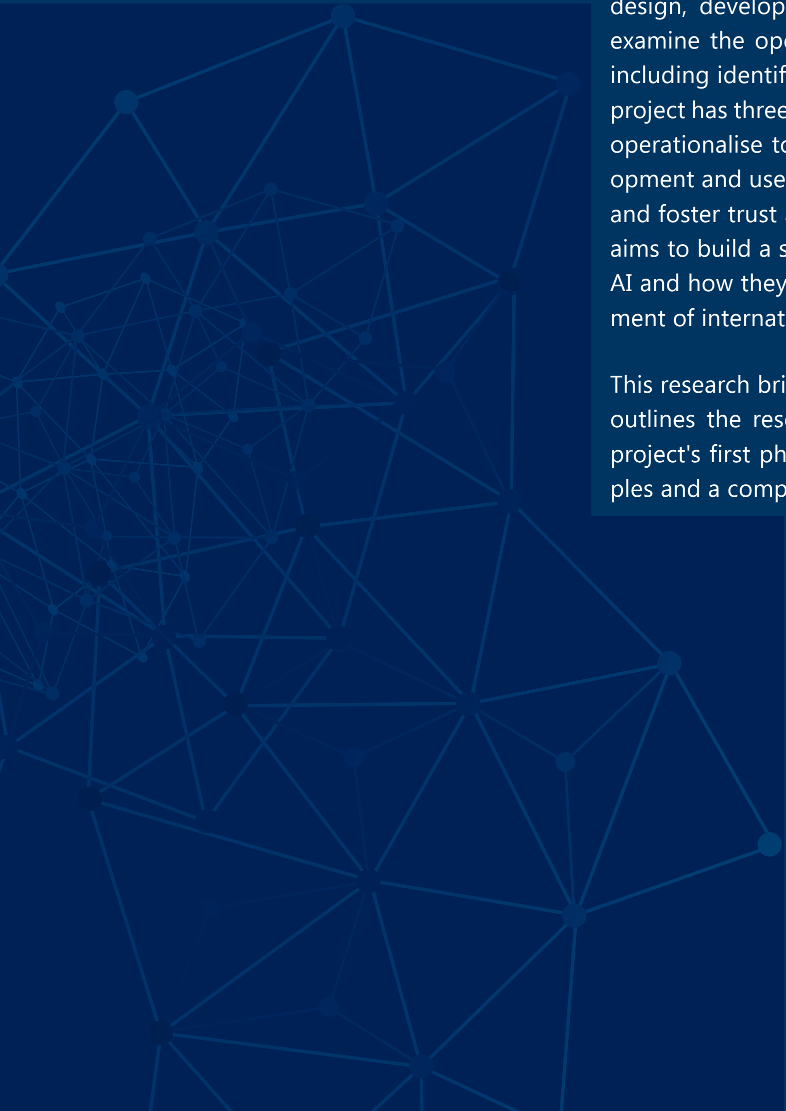
AI	Artificial Intelligence
GGE	Group of Governmental Experts
LAWS	Lethal Autonomous Weapons Systems
NATO	North Atlantic Treaty Organization
WEOG	Western European and Others Group

EXECUTIVE SUMMARY

Continuous advances in the field of artificial intelligence (AI) and efforts to integrate AI systems in critical sectors are gradually transforming all aspects of society, including in the defence sector. Although advancements in AI present unprecedented opportunities to augment human capabilities and improve decision-making in various ways, they also present significant legal, safety, security and ethical concerns. Thus, to ensure that AI systems are developed and used lawfully, ethically, safely, securely and responsibly, governments and intergovernmental organisations are developing a range of normative instruments. This approach is broadly known as "Responsible AI", or ethical or trustworthy AI. Presently, the most notable approach to Responsible AI is the development and operationalisation of responsible or ethical AI principles.

UNIDIR's project **Towards Responsible AI in Defence** seeks to, first, build a common understanding of the key facets of responsible research, design, development, deployment, and use of AI systems. It will then examine the operationalisation of Responsible AI in the defence sector, including identifying and facilitating the exchange of good practices. The project has three main aims. First, it aims to encourage states to adopt and operationalise tools that can enable responsible behaviour in the development and use of AI systems. It also seeks to help increase transparency and foster trust among states and other key AI actors. Finally, the project aims to build a shared understanding of the key elements of Responsible AI and how they may be operationalised, which may inform the development of internationally accepted governance frameworks.

This research brief provides an overview of the aims of the project. It also outlines the research methodology for and preliminary findings of the project's first phase: the development of a common taxonomy of principles and a comparative analysis of AI principles adopted by states.



ABOUT THE PROJECT

Advances in the field of AI and efforts to integrate AI systems in critical sectors are gradually transforming all aspects of our society – the defence sector is no exception.[1] AI developments and their applications present unprecedented opportunities to augment human capabilities and improve decision-making in various ways, particularly in problem-solving, data processing and decision-making. However, significant legal, safety, security and ethical concerns relating to AI adoption are coming to light as AI systems are increasingly being deployed worldwide across sectors. These concerns include issues related to transparency, reliability, predictability, understandability, accountability, bias and discrimination, and technical robustness. Such concerns are heightened in certain high-risk military contexts, where errors or misuses could result in serious injury, loss of life or damage to critical infrastructure.

It is therefore essential that AI systems are developed and used in a responsible and safe manner and in accordance with legal requirements and ethical values. To ensure this, governments and intergovernmental organisations along with private actors are developing and adopting a range of governance instruments – such as principles, standards and codes of conduct – to guide AI research, design, development, deployment and use across sectors.[2] This varied approach to AI governance is broadly known as ethical or trustworthy AI or, as here, "Responsible AI". At present, Responsible AI initiatives often begin with the adoption of AI principles that articulate the requirements that AI systems should meet so that they can be used lawfully, ethically, safely, securely and responsibly.

However, Responsible AI is an emerging and evolving field of research and practice, particularly in the defence sector. Only a handful of states and intergovernmental organisations have publicly adopted principles, standards or ethical frameworks tailored to AI applications in the defence sector.

While many initiatives to map and assess the operationalisation of AI principles have started to emerge, there is a need for more dedicated work to map and analyse AI principles developed by states, especially with respect to their application in the defence sector. Such an exercise could have three positive outcomes. First, it could encourage states to adopt and operationalise tools that can enable responsible behaviour in the development and use of AI. It could also help to increase transparency and foster trust among states and other key AI actors.[3] Finally, it could build a shared understanding of the key elements of Responsible AI and how they may be operationalised, which may inform the development of internationally accepted governance frameworks.

The UNIDIR project **Towards Responsible AI in Defence** is a step in this direction. It aims to build a common understanding of the key facets of responsible research, design, development, deployment and use of AI systems. Further, it aims to examine the operationalisation of Responsible AI in the defence sector, including identifying and facilitating the exchange of good practices. To do so, it includes two phases. In the first phase, through desk research and analysis of existing AI principles adopted by states and intergovernmental organisations, it aims to develop a common taxonomy of AI principles and to identify commonalities in states' views on the essential elements of Responsible AI. In the second phase, through stakeholder interviews and workshops, the project seeks to examine how AI principles are and could be operationalised in the defence domain; what structures may need to be put in place for their operationalisation; and the associated gaps and challenges and how they may be addressed.

Through the two phases, the project will address the following research questions:

Phase 1

- Which states and intergovernmental organisations have adopted Responsible AI principles exclusively for the defence sector or have national AI principles whose stated application may extend to the defence sector?

[1] For the definition of "AI systems", see OECD (n.d.).

[2] For the purpose of this paper, the "AI system life cycle" refers to the range of activities from "research, design and development to deployment and use" of an AI system. See UNESCO (2022, 4).

[3] For the purpose of this paper, "AI actors" refer to "any actor involved in at least one stage of the AI system life cycle, and can refer both to natural and legal persons, such as researchers, programmers, engineers, data scientists, end-users, business enterprises, universities and public and private entities, among others". See UNESCO (2022, 1).

- What are the essential facets or principles of responsible, ethical or trustworthy research, design, development, deployment and use of AI systems? What do these principles refer to?
- Which of these principles do states commonly consider as being essential?

Phase 2

- What does Responsible AI refer to in the defence domain? What types of tools are needed to ensure the responsible, ethical or trustworthy research, design, development, deployment and use of AI systems in defence?
- What are the existing tools for Responsible AI that are tailored or applicable to the defence sector? How are they being operationalised? What are the commonalities and differences in national approaches to operationalising Responsible AI?
- What governance structures need to be put in place for the operationalisation of Responsible AI?
- What are the challenges, if any, to the operationalisation of Responsible AI? How can they be addressed?
- How are states ensuring that the civilian AI industry that works with defence organisations aligns with their Responsible AI standards and requirements?
- What should Responsible AI education and training entail and to whom should it be provided?

This research brief provides an overview of the methodology and key preliminary findings of the first phase. This stage involves the identification of AI principles adopted by states and intergovernmental organisations and the building of a common taxonomy based on the identified AI principles. A comparative analysis of the identified principles against the common taxonomy then reveals the commonalities and differences in states' perspectives on essential elements of Responsible AI. However, this is a "living" research project that will continue to develop as the global AI policy landscape evolves, particularly for the defence sector. Therefore, the common taxonomy is not intended to be authoritative and will be updated as required.

METHODOLOGY FOR PHASE 1

To conduct a comparative analysis of AI principles adopted by states, a common, sector-agnostic taxonomy of AI principles was developed. The common taxonomy comprises a list of AI principles with brief definitions for each. The definitions are brief as they embody the "lowest common denominator" understanding of what that principle refers to among those states and intergovernmental organisations that have adopted it. For this reason, the common taxonomy serves as a tool against which AI principles can be compared in order to identify which principles are most and least commonly adopted by states to ensure Responsible AI.

The reason for a sector-agnostic taxonomy is twofold. First, in practical terms, only a few states and intergovernmental organisations have adopted defence-specific AI principles. This is insufficient to build a common taxonomy tailored exclusively to the defence sector. From a conceptual point of view, given the paucity of defence-specific principles, a sector-agnostic taxonomy could still be relevant for the analysis as AI is a general-purpose technology. As such, guidance to ensure AI systems are researched, designed, developed, deployed and used in a responsible, legal, ethical, safe and secure manner may be generally relevant across sectors, although there will be differences in the way they are operationalised. Moreover, since defence organisations are working with the civilian AI industry to build AI systems, AI principles developed for civilian sectors may also be relevant for the defence sector.

STEP 1. MAPPING AND ANALYSING AI PRINCIPLES ADOPTED BY INTERGOVERNMENTAL ORGANISATIONS

The first step in developing the common taxonomy is to map AI principles adopted by intergovernmental organisations. AI principles adopted by intergovernmental organisations are used as the basis for the common taxonomy because they are arguably the most explicit embodiment of shared understandings on what attributes and requirements constitute Responsible AI.

In this initial stage, extensive desk-based research identified AI principles adopted by intergovernmental organisations. Eleven intergovernmental organisations that have each developed and adopted some variation of AI principles were identified (listed in Annex B). The AI principles developed and adopted by the 11 intergovernmental organisations are in line with the mandate of the respective organisation and thus differ in scope – some are applicable across sectors, while others are sector-specific. Only two are tailored specifically to the defence domain – the North Atlantic Treaty Organization (NATO) *Principles of Responsible Use of Artificial Intelligence in Defence* and the *Guiding Principles* adopted by the United Nations Group of Governmental Experts (GGE) on lethal autonomous weapons systems (LAWS) to guide the future work of the GGE and to provide a framework for the development and use of LAWS.[4] Through a careful study of the 11 sets of AI principles, a base list of broad AI principles was developed.[5] These principles either featured explicitly as a stand-alone principle or were reflected implicitly in the explanations within the sets of principles. Where necessary, some of these broad principles were further subdivided into separate, narrower principles to capture different aspects and nuances (see Table 1).[6]

[4] It is important to note that, while the Guiding Principles adopted by the GGE on LAWS specifically focus on autonomy in weapon systems, the NATO Principles are broader in scope. They concern a range of military uses of AI beyond weapon systems.

[5] Note that some of the principles may not apply in the defence sector or may apply differently.

[6] Note that these are preliminary definitions. See Table 2 below for the final common taxonomy and definitions.

TABLE 1. AI PRINCIPLES ADOPTED BY INTERGOVERNMENTAL ORGANISATIONS[7]

PRINCIPLE		DEFINITION
Fairness	**Impartiality**	AI systems should not follow or create biases or discrimination
	Inclusiveness	Access to the benefits of AI systems should be equal
	Human Oversight, Judgement or Control	Users should be informed actors and exercise appropriate levels of oversight, judgement or control over AI systems, including the ability to avoid unintended consequences
	Human Dignity	AI systems should not violate the inherent human state of being worthy of respect
Lawfulness	**Compliance with Law**	All AI systems should be developed and used in accordance with national and international law
	Data Protection	AI systems should only use data where informed consent has been given, if appropriate
	Privacy	AI systems should not interfere with the right to private life
	Proportionality	The AI system chosen should be proportional to achieve a given legitimate objective
	Public Engagement	Collaborations and exchanges with stakeholders should be encouraged for the development and governance of AI
Responsibility and Accountability	**Accountability**	Accountability should always be attributed to the appropriate human actors
	Responsibility	AI systems should be used under the appropriate conditions and by appropriately trained individuals, who remain responsible
	Sustainability	AI systems should yield beneficial outcomes for people and the planet, where applicable
Technical Robustness	**Reliability**	AI systems should be tested appropriately to ensure that they function as intended in the circumstances of their use
	Safety	Unintended harms should be avoided, prevented and addressed
	Security	Vulnerabilities to adversarial attacks should be addressed, prevented and eliminated to the extent possible
Transparency	**Explainability**	Users should be able to appropriately understand the outcomes of AI systems and how conclusions are reached
	Information sharing	Users should be informed when an AI system is used
	Traceability	Data sets, processes and decisions of AI systems should be made open to analysis and inquiry, if appropriate

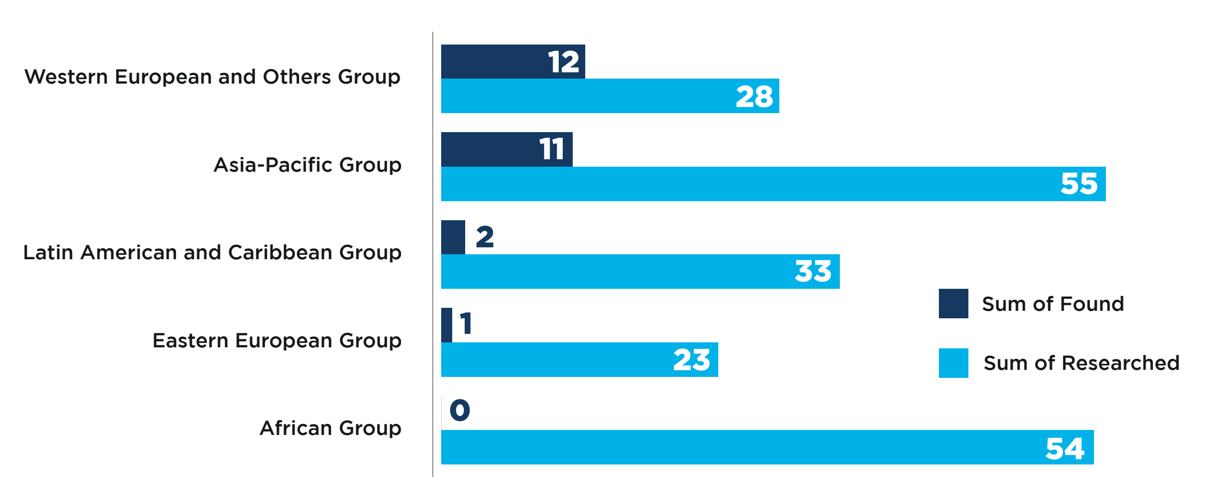
[7] Ibid.

STEP 2: MAPPING AND ANALYSING AI PRINCIPLES ADOPTED BY STATES

The second stage is to map AI principles adopted by states. First, AI principles developed exclusively for application in the defence sector were identified. As very few states currently have publicly adopted defence-specific principles, national AI principles adopted by states were included when their stated application does not exclude the defence sector.

Of the 193 states researched, **26 states were found to have adopted a set of AI principles** (listed in Annex B). The 26 states include 12 from the Western European and Others Group (WEOG), 11 from the Asia-Pacific Group, two from the Latin American and Caribbean Group and one from the Eastern European Group.[8]

FIGURE 1. NUMBER OF STATES THAT HAVE ADOPTED AI PRINCIPLES BY REGIONAL GROUP



Only two states – the United States and the United Kingdom – have adopted a set of AI principles specifically for defence. Three states – France, Australia and Canada – that have not yet adopted a set of principles for defence have instead developed other instruments to guide the development and use of AI in the defence domain, such as road maps, ethical risk guidelines and assessment frameworks (see Annex B).[9] As the analysis requires a 1:1 comparison of principles, it does not include the defence-specific AI governance frameworks adopted by France, Australia and Canada. They will be examined in the second phase of the project, which will focus on the operationalisation of Responsible AI instruments in defence.

Using the same methodology adopted for intergovernmental organisations, the AI principles adopted by states were analysed to **build upon and refine the preliminary common taxonomy to ensure that it reflects the principles adopted by states and what they commonly understand those principles to entail.** In doing so, additional principles were identified that featured more prominently in principles adopted by states. These include international cooperation, risk-based approach, human autonomy and should not harm.[10] At the same time, the principle of proportionality was merged with the principle of compliance with law because many of the states that included proportionality in their principles discussed it in context of international law.

Furthermore, three different interpretations of the principle of reliability were found. The first, termed

[8] See United Nations (n.d.).

[9] French Ministry of Armed Forces (2019); Devitt et al. (2021); Defence Research and Development Canada (2017).

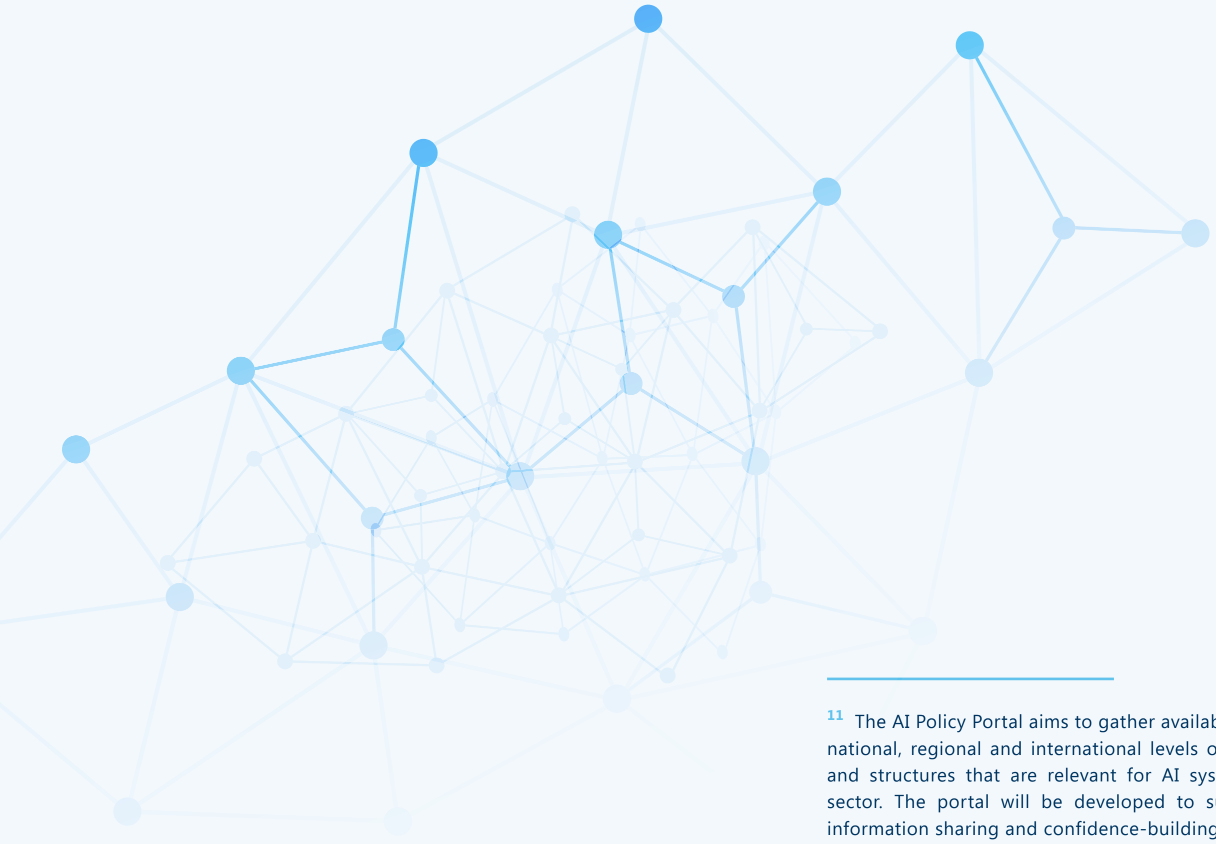
[10] For definitions of these principles, see Table 2.

resilience, is that AI systems should be tested appropriately to ensure that they function as intended in the circumstances of their use. The second, termed redundancy, is that AI actors and users should not be over-reliant on AI systems and should possess the ability to continue operations as appropriate in case of failure of AI system(s). The third, termed data quality, is that AI actors should ensure that AI systems are trained on data of sufficient quality, remove corrupt data and have quality control checks (either ex-ante or ex-poste) to ensure reliable and valid results.

All the principles identified in the two-step process described above form the common taxonomy (see Table 2 and Figure 2). This forms the basis for the comparative analysis of AI principles adopted by states.

IMPORTANT CAVEATS CONCERNING THE METHODOLOGY AND ANALYSIS

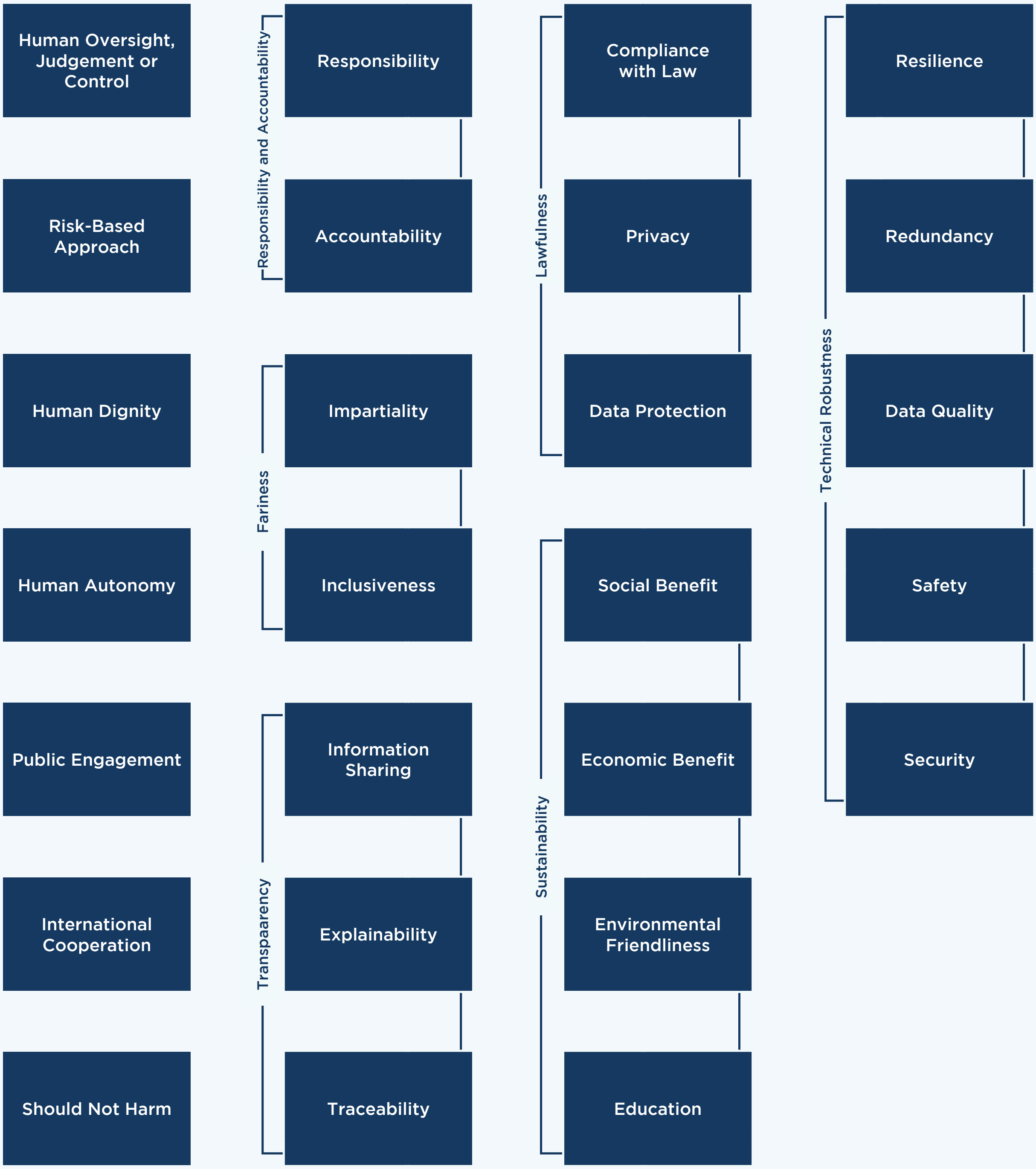
- The analysis focuses on a 1:1 comparison of principles. It therefore only considers officially adopted set of AI principles. It does not include states that have an AI strategy or similar documents but have not officially adopted a set of principles.
- As the analysis includes AI principles that are not tailored exclusively to the defence sector, some of the principles in the common taxonomy may not apply in the defence sector or may apply differently.
- The findings are based on desk research. There may be other states that have AI principles that could not be identified through desk research, for example due to translation issues. The common taxonomy and comparative analysis will be built on and continually refined based on the data collected by UNIDIR's upcoming AI Policy Portal.[11] The findings are not exhaustive.
- This is a living research project that will develop as the global AI policy landscape evolves, particularly in the defence sector. Therefore, the common taxonomy is not intended to be authoritative. Rather, it serves as a tool used to comparatively analyse AI principles adopted by states.



[11] The AI Policy Portal aims to gather available information at the national, regional and international levels on policies, processes and structures that are relevant for AI systems in the defence sector. The portal will be developed to support transparency, information sharing and confidence-building in the field of AI.

THE COMMON TAXONOMY

FIGURE 2. CATEGORISATION OF THE COMMON TAXONOMY



Notes: *The Figure shows the principles (in blue boxes) that form the common taxonomy. It also demonstrates which broad principles (in brackets) were subdivided into "narrower" principles to capture their different aspects and nuances. Only the "narrow" principles are included in the common taxonomy as explained in the methodology. Additionally, some of the principles may not apply in the defence sector or may apply differently. See Table 2 for definitions of the principles. Lastly, the principles in the figure have been arranged for visual ease, rather than in order of priority.*

TABLE 2. THE COMMON TAXONOMY

PRINCIPLE	DEFINITION
Human Oversight, Judgement or Control	AI actors should be informed actors and should exercise appropriate levels of oversight, judgement or control of the choices made – whether, when and how to delegate decisions and actions to AI systems; the ability to detect and avoid unintended consequences; and the ability to take steps (e.g. disengagement, provision of recourse or deactivation of systems) when such systems demonstrate unintended behaviour
Risk-Based Approach	A preventative approach should be taken to minimise negative impacts and ensure that AI systems are used to achieve the intended goal through anticipating potential risks, taking measures to minimise those risks and taking mitigating actions to avert unintended harms
Human Dignity	AI systems should not violate the inherent human state of being worthy of respect
Human Autonomy	AI system should preserve human autonomy so that AI actors are able to make independent and informed decisions without AI systems removing their self-determination
Public Engagement	There should be open collaboration and exchanges with stakeholders for the development and governance of AI
International Cooperation	International cooperation should be pursued in order to avoid a malicious arms race and to promote safety regulations
Should Not Harm	AI systems should not be developed with the aim of harming or deceiving humans
Responsibility	AI actors should ensure that AI systems are developed and used with appropriate levels of human judgement and care and by suitably trained human actors who remain responsible
Accountability	Human actors should remain accountable for the decisions and actions performed by or based on an AI system, in accordance with their role in the AI system's life cycle
Impartiality	AI systems should not create, follow or reinforce unintended biases
Inclusiveness	All should be able to access the benefits of AI
Information Sharing	AI actors should be duly informed when a decision is assisted by or made by an AI system or when their data is being collected for use in an AI system, if appropriate
Explainability	Relevant AI actors should be able to appropriately understand the outcome of an AI system and/or how the system arrived at its outcome
Traceability	Processes involved in AI systems should be documented to enable analysis of the AI system's outcomes and to address inquiries and audits, if appropriate
Compliance with Law	All AI systems should be researched, designed, developed, deployed, and used in compliance with national and international law
Privacy	AI systems should not infringe on the right to private life
Data Protection	AI systems should ensure that the collection, use and disposal of personal data respects appropriate national and international data protection regulations where relevant, such as when training AI algorithms

Social Benefit	AI systems should generate measurable social benefits, such as increase in quality of life or public well-being, promote inclusive development and universal welfare, narrow disparities, improve safety, and avoid a malicious AI race
Economic Benefit	AI systems should generate measurable economic benefits, such as increasing or facilitating economic competitiveness, creating innovations, facilitating new industries, adding value to economic processes, improving the economic well-being of people and economic stability, and helping the economic adaptability of job losses resulting from AI innovations
Environmental Friendliness	The development of AI systems should take place in an environmentally friendly manner, including ensuring the protection of the environment and environmental resources
Education	AI-relevant training and educational programmes based on the latest developments should be provided at the societal scale to ensure responsible development and use of AI systems and prevent unintended harms or malicious use and generating or exacerbating inequalities
Resilience	AI systems should be tested appropriately to ensure that they function as intended in the circumstances of their use
Redundancy	AI actors and users should not be over-reliant on AI systems and should possess the ability to continue operations as appropriate in case of failure of AI system(s)
Data Quality	AI actors should ensure that AI systems are trained on data of sufficient quality, remove corrupt data and have quality control checks (either ex-ante or ex-poste) to ensure reliable and valid results
Safety	Unintended harms should be avoided, addressed, prevented and eliminated throughout the life cycle of an AI system
Security	Vulnerabilities to attack should be avoided, addressed, prevented and eliminated throughout the life cycle of an AI system

COMPARATIVE ANALYSIS OF AI PRINCIPLES ADOPTED BY STATES

To identify a range of principles that states commonly perceive as being essential elements of Responsible AI, a comparative analysis of the AI principles adopted by states was conducted based on the common taxonomy. To conduct the analysis, each state's AI principles were analysed and compared against the common taxonomy to identify which principles in the taxonomy feature among the principles adopted by the state (see annex A for numbers of states that have adopted each principle in the common taxonomy). The analysis considered both explicitly adopted, stand-alone principles as well as those that were implicitly referred to in the explanations of the stand-alone principles. A comparison of the AI principles adopted by states against the common taxonomy revealed that the most commonly adopted principles include impartiality, inclusiveness, safety, human oversight, judgement or control, compliance with law, responsibility and social benefit (see Figure 3). In contrast, the least commonly adopted principles include redundancy, international cooperation, education, should not harm, public engagement, environmental friendliness, data quality, human autonomy and economic benefit (see Figure 4).

FIGURE 3. MOST COMMONLY ADOPTED AI PRINCIPLES

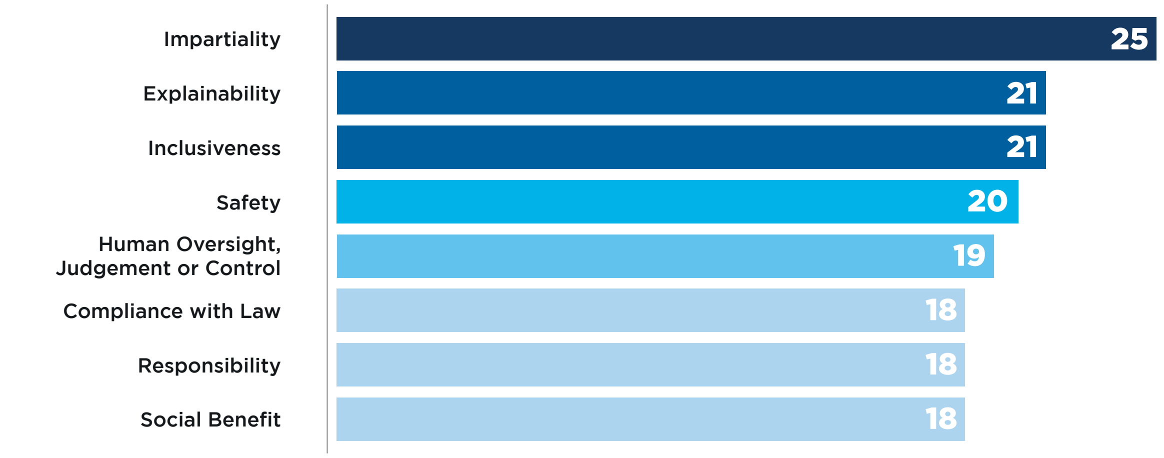
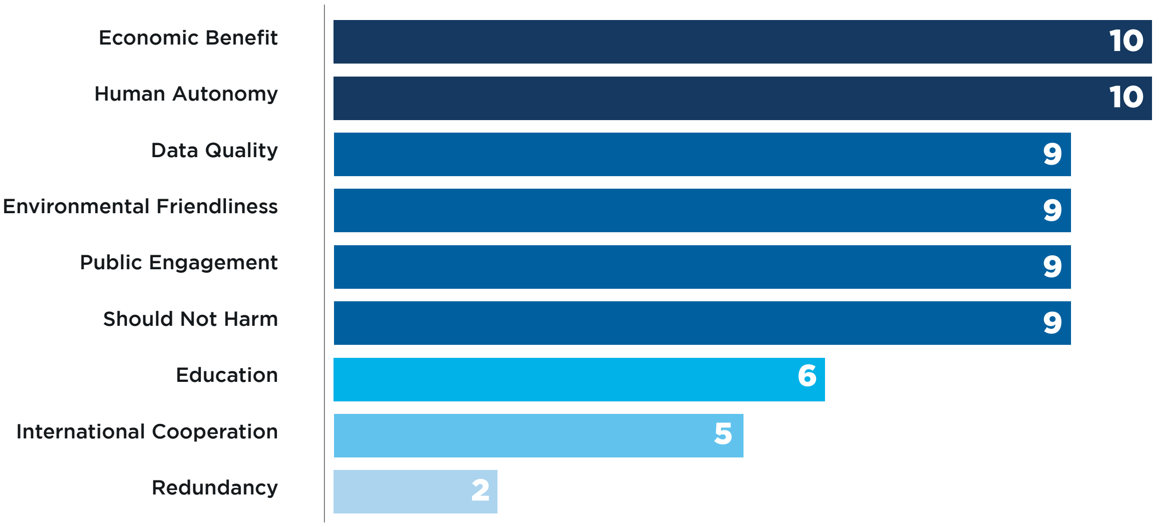


FIGURE 4. LEAST COMMONLY ADOPTED AI PRINCIPLES



Disaggregating the data between explicitly and implicitly adopted principles shows which principles were most commonly adopted as stand-alone principles and which are most commonly adopted implicitly in the definition of another principle. In general, the principles adopted explicitly tend to focus on the technical characteristics of AI systems, for example, impartiality, explainability and human oversight, judgement or control were the most common explicitly adopted principles (see Figure 5). In contrast, principles adopted implicitly tend to focus on the societal impact of AI systems, such as public engagement and social benefit (see Figure 6).

FIGURE 5. MOST EXPLICITLY ADOPTED AI PRINCIPLES

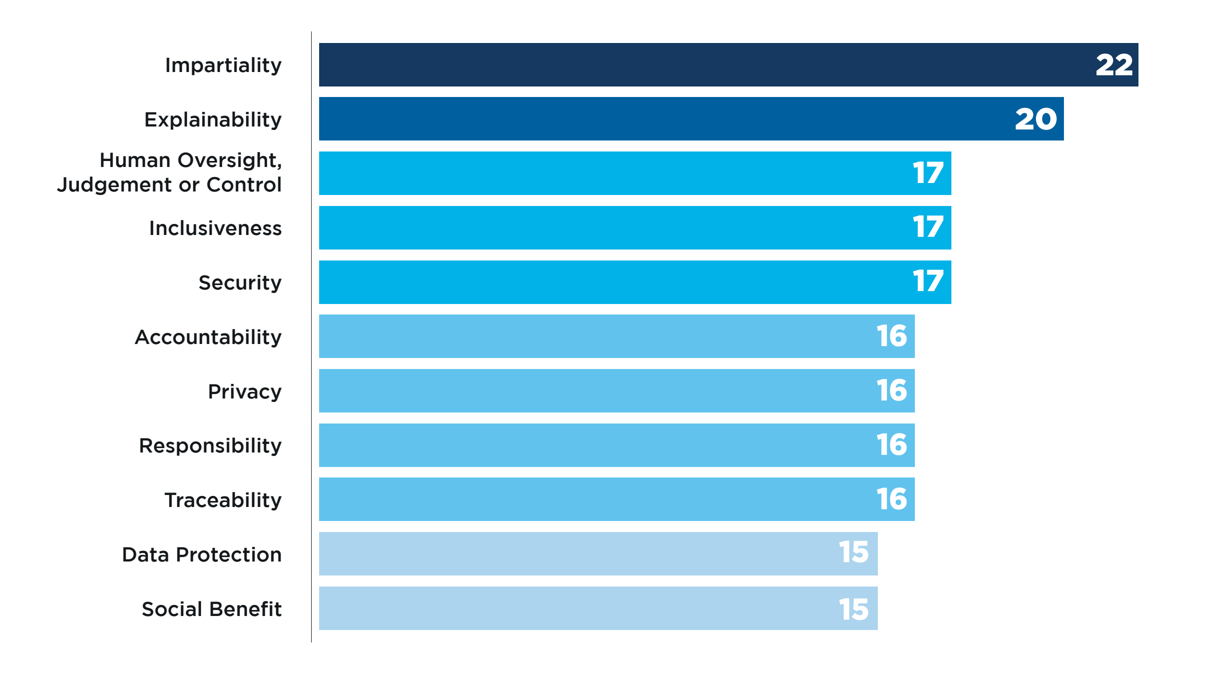
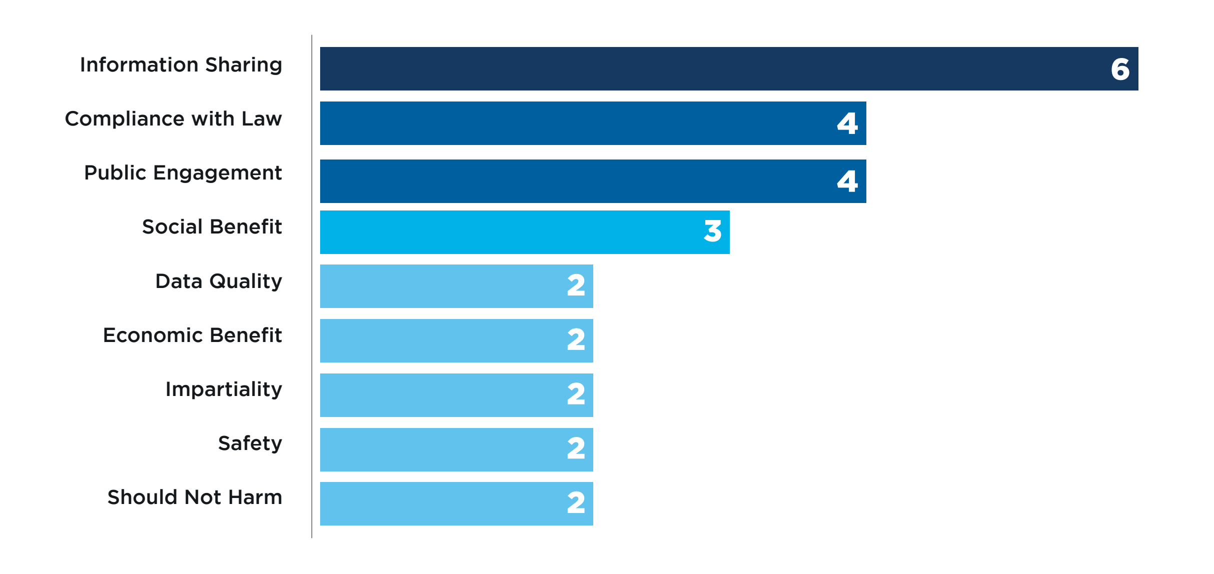


FIGURE 6. MOST IMPLICITLY ADOPTED AI PRINCIPLES



Furthermore, while there are not enough data points to conduct a comprehensive geographical comparative analysis, it was possible to conduct a preliminary geographical analysis for the Asia-Pacific Group and Western European and Others Group. When comparing these two regions, there were some similarities: the five most adopted principles of both groups include safety as well as impartiality and inclusiveness – both of which fall under the broad principle of fairness. In contrast, social benefit and risk-based approach featured more commonly in the case of the Asia-Pacific Group than explainability and compliance with law, which were more prominent in the Western European and Others Group.

FIGURE 7. MOST COMMONLY ADOPTED AI PRINCIPLES, WESTERN EUROPEAN AND OTHERS GROUP

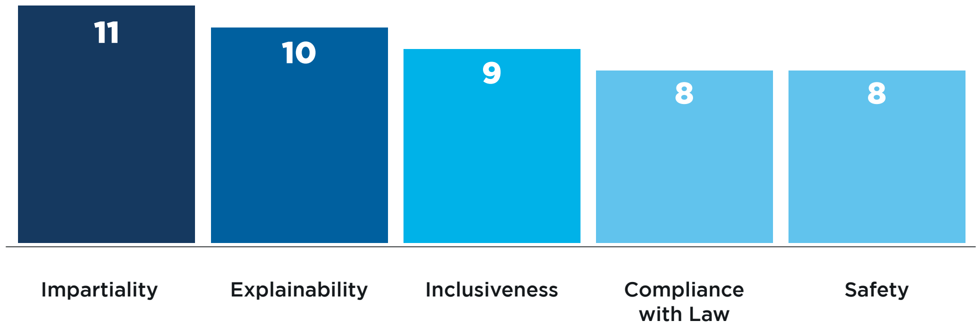
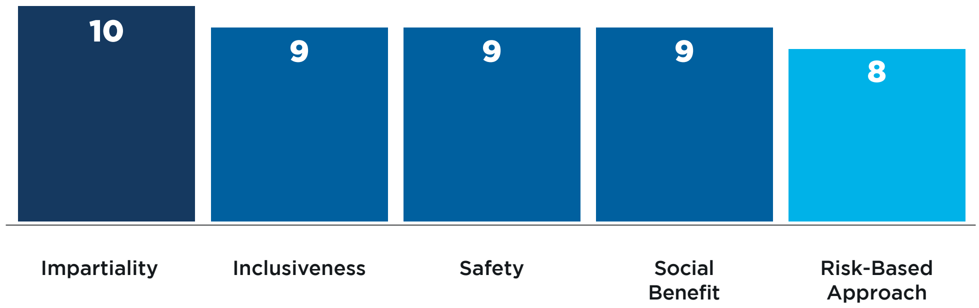


FIGURE 8. MOST COMMONLY ADOPTED AI PRINCIPLES, ASIA-PACIFIC GROUP



CONCLUSION AND NEXT STEPS

The research shows that, while the global AI policy landscape is still at a nascent stage, it is gradually evolving. An increasing number of states are developing AI strategies and policy instruments to guide the lawful, ethical, safe, secure and responsible research, design, development, deployment and use of AI. Among those that have adopted principles to this end there are commonalities in terms of which principles should inform the different stages of the AI life cycle from ideation to use, and a degree of shared understanding on what those principles entail.

In the defence sector, currently only a handful of states and intergovernmental organisations have developed AI principles exclusively to guide AI applications in the defence sector. While in the next few years more states can be expected to adopt principles, the research presented here shows that states may have different approaches to ensuring the responsible use of AI in defence. Some states may not necessarily adopt defence AI principles. Instead, some states may develop other Responsible AI instruments such as ethical AI risk-assessment frameworks, codes of conduct or a combination of instruments. Some may also use their sector-agnostic national AI principles or technology-agnostic ethics guidelines to guide the research, design, development, deployment and use of AI for defence purposes.

Regardless of which Responsible AI tool is adopted, it is essential that it is put into practice in an effective and continuous manner. To this end, the second phase of this project will study existing defence-specific Responsible AI instruments and will explore the kinds of Responsible AI tools that are suitable for the defence sector. It will examine how different Responsible AI tools can be operationalised, what gaps and challenges there are to effective operationalisation, and how they may be addressed. The project will also address what new governance structures, if any, need to be put in place for the operationalisation of Responsible AI instruments in the defence sector. Moreover, since defence organisations work with the civilian AI industry to build AI systems, the next phase of the project will explore what measures can be taken to ensure that the civilian AI industry that works with defence organisations aligns with Responsible AI standards and requirements.

ANNEX A. FREQUENCY OF ADOPTION OF AI PRINCIPLES

Number of states that have adopted each principle in the Common Taxonomy

PRINCIPLE	NUMBER OF STATES
Impartiality	25
Explainability	21
Inclusiveness	21
Safety	20
Human oversight, judgment or control	19
Security	19
Compliance with law	18
Responsibility	18
Social benefit	18
Accountability	17
Information sharing	17
Privacy	17
Traceability	17
Data protection	16
Human dignity	12
Resilience	12
Risk-based approach	11
Economic benefit	10
Human autonomy	10
Data quality	9
Environmental friendliness	9
Public engagement	9
Should not harm	9
Education	6
International cooperation	5
Redundancy	2

ANNEX B. DATA SOURCES

The tables below include links to AI Principles documents (adopted by states and intergovernmental organisations) examined in this project.

INTERGOVERNMENTAL ORGANISATIONS

ORGANISATION	DOCUMENT
North Atlantic Treaty Organization (NATO)	NATO Principles of Responsible Use of Artificial Intelligence in Defence
European Union (EU)	Ethics Guidelines for Trustworthy AI
United Nations Educational, Scientific and Cultural Organization (UNESCO)	Recommendation on the Ethics of Artificial Intelligence
Group of Government Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (GGE on LAWS)	Guiding Principles affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System
Council of Europe	European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment
Organisation for Economic Co-operation and Development (OECD)	OECD Recommendation of the Council on Artificial Intelligence
International Finance Corporation (IFC)	IFC Technology Code of Conduct – Progression Matrix
World Health Organization (WHO)	Key Ethical Principles for Use of AI for Health
Global Privacy Assembly (GPA)[12]	Declaration on Ethics and Data Protection in Artificial Intelligence
Food and Agriculture Organization (FAO), Microsoft, IBM, Italian National Ministry of Innovation, Pontifical Academy for Life and others	Rome Call for AI Ethics
Digital Economy Partnership Agreement (DEPA)	Module 4 (Data Issues) and Module 8 (Emerging Trends and Technologies) of the Agreement

[12] The GPA was previously the International Conference of Data Protection and Privacy Commissioners.

STATES

STATE	DOCUMENTS
Asia-Pacific Group	
China	• Governance Principles for the New Generation Artificial Intelligence • Beijing Artificial Intelligence Principles • Joint Pledge on Artificial Intelligence Industry Self-Discipline (Draft for Comment)
Cyprus	• National AI Strategy
India	• Responsible AI – Approach Document for India
Indonesia	• Strategi Nasional Kecerdasan Artifisial
Japan	• AI R&D Guidelines for International Discussions • Social Principles of Human-Centric AI
Jordan	• Jordan Artificial Intelligence Policy
Republic of Korea	• National Strategy for Artificial Intelligence • Artificial Intelligence Personal Information Protection Self-Checklist
Singapore	• Model Artificial Intelligence Governance Framework – 2nd Edition
Thailand	• Digital Thailand AI Ethics Guidelines
Türkiye	• National Artificial Intelligence Strategy 2021–2025
United Arab Emirates	• Smart Dubai AI Ethics Principles and Guidelines
Eastern European Group	
Russian Federation[13]	• The Code of Ethics in the Field of Artificial Intelligence
Ukraine	• National Strategy for Development of Artificial Intelligence
Latin American and Caribbean Group	
Colombia	• Marco Ético para la Inteligencia Artificial en Colombia
Uruguay	• Estrategia de Inteligencia Artificial para el Gobierno Digital
Western European and Others Group	
Australia	• Australia's AI Ethics Principles • A Method for Ethical AI in Defence
Austria	• Artificial Intelligence Mission Austria 2030 – AIM AT 2030

[13] Russia's Code of Ethics precludes it from application in the military context and therefore it was not included in the analysis. However, the Code of Ethics was studied as the common taxonomy was built.

Canada	• Responsible Use of AI – Guiding Principles • Directive on Automated Decision-Making • A Framework to Assess the Military Ethics of Human Enhancement Technologies • A Framework to Assess the Military Ethics of Emerging Technologies
Denmark	• National Strategy for Artificial Intelligence
Finland	• Work in the Age of Artificial Intelligence: Four Perspectives on the Economy, Employment, Skills and Ethics
France[14]	• For a Meaningful Artificial Intelligence: Towards a French and European Strategy • Artificial Intelligence in Support of Defence • Defence Ethics Committee: Opinion on the Integration of Autonomy into Lethal Weapon Systems
Israel	• Harnessing Innovation: Israeli Perspectives on AI Ethics and Governance
Malta	• The Ultimate AI Launchpad: A Strategy and Vision for Artificial Intelligence in Malta 2030
Norway	• National Strategy for Artificial Intelligence
Portugal	• AI Portugal 2030
Switzerland	• Guidelines for Artificial Intelligence for the Confederation
United Kingdom	• Defence Artificial Intelligence Strategy • Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-Enabled Capability in Defence
United States	• Department of Defense Adopts Ethical Principles for Artificial Intelligence

[14] The documents listed here concern France's approach to AI in defence. However, since France has not yet adopted a set of principles these documents were studied, but not included in the analysis.

REFERENCES

Defence Research and Development Canada. 2017. "A framework to assess the military ethics of human enhancement technologies". As of 9 January 2023: https://cradpdf.drdc-rddc.gc.ca/PDFS/unc279/p805510_A1b.pdf.

Devitt, Kate et al. 2021. "A method for ethical AI in defence". As of 9 January 2023: https://www.dst.defence.gov.au/publication/ethical-ai.

French Ministry of Armed Forces. 2019. "Artificial intelligence in support of defence". AI Task Force. As of 9 January 2023: https://www.defense.gouv.fr/sites/default/files/aid/Report%20of%20the%20AI%20Task%20Force%20September%202019.pdf.

Organisation for Economic Co-operation and Development (OECD). n.d. "AI terms and concepts". As of 6 January 2023: https://oecd.ai/en/ai-principles.

United Nations. n.d. "Regional groups of Member States". As of 6 January 2023. https://www.un.org/dgacm/en/content/regional-groups.

United Nations Educational, Scientific and Cultural Organization (UNESCO). 2022. "Recommendations on the ethics of artificial intelligence". As of 6 January 2023: https://unesdoc.unesco.org/ark:/48223/pf0000381137.

UNIDIR
UNITED NATIONS INSTITUTE
FOR DISARMAMENT RESEARCH

@unidirgeneva

@UNIDIR

@un_disarmresearch