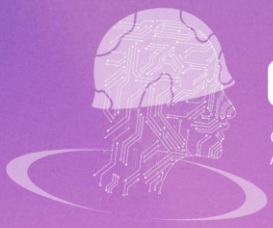

GC REAIM

GLOBAL COMMISSION ON RESPONSIBLE AI IN THE MILITARY DOMAIN

Foreign, Commonwealth & Development Office

GC REAIM Expert Policy Note Series

Principles of Responsible Military Artificial Intelligence and Applicable International Law

Rain Liivoja

May 2025

POWERED BY

The Hague Centre for Strategic Studies

GC REAIM Expert Policy Note Series

Principles of Responsible Military Artificial Intelligence and Applicable International Law

Authors: Rain Liivoja

May 2025

Cover photo: Unsplash

The Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM) is an initiative of the Government of the Netherlands that was launched during the 2023 REAIM Summit on Responsible Artificial Intelligence in the Military Domain in The Hague. Upon request of the Dutch Ministry of Foreign Affairs, the Hague Centre for Strategic Studies acts as the Secretariat of the Commission.

The GC REAIM Expert Policy Note Series was funded by the Foreign, Commonwealth and Development Office (FCDO) of the United Kingdom. GC REAIM Experts maintained full discretion over the topics covered by the Policy Notes. The contents of the GC REAIM Expert Policy Note series do not represent the views of the Global Commission as a whole. The Policy Notes are intended to highlight key issues related to the governance of AI in the military domain and provide policy recommendations.

HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
@hcssnl

The Hague Centre for Strategic Studies
Email: info@hcss.nl
Website: www.hcss.nl

1. Introduction

The strong interest of armed forces in artificial intelligence (AI) has raised concerns about the adequacy of existing policies, standards, and rules, especially as regards the protection of civilians in armed conflict. The absence of a tailor-made legal framework and uncertainties about the interpretation of existing law have led States to draft and promulgate guiding principles on the development and use of AI for military purposes. Some of these principles are now being considered for inclusion by reference in an instrument—potentially a legally-binding instrument—developed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems ('GGE').[1]

This raises questions about the normative significance of these principles and their interrelationship with the existing legal framework. This paper first identifies several sets of such principles for comparison. It then examines whether any of the specific principles correlate to existing rules and principles of international law applicable in armed conflict, or whether they add to the existing governance framework by interpreting the law or building upon it. The paper offers three conclusions regarding the inclusions of the principles in a potential future instrument, especially one that is legally binding.

[1] Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE LAWS), *Revised Rolling Text as of 8 November 2024*, Convention on Certain Conventional Weapons, United Nations Office for Disarmament Affairs, 2024, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/Revised_rolling_text_as_of_8_November_2024_final.pdf; See also Argentina, Ecuador, El Salvador, Colombia, Costa Rica, Guatemala, Kazakhstan, Nigeria, Palestine, Panama, Peru, Philippines, Sierra Leone and Uruguay, 'Draft Protocol on Autonomous Weapon Systems (Protocol VI)' (11 May 2023) CCW/GGE.1/2023/WP.6, art 4.

2. The Proliferation of AI Governance Principles

Various sets of guiding principles on military AI have emerged from work undertaken on the national level, as well as through several multilateral processes. Among domestic initiatives seeking to provide guidance to national military establishments and defence industries, the following are some of the more prominent examples:

- In 2020, the United States Department of Defence adopted five 'Ethical Principles for Artificial Intelligence'.[2]
- In 2021, Singapore's Ministry of Defence established four 'preliminary AI guiding principles'.[3]
- In 2022, the United Kingdom Ministry of Defence published a policy paper titled 'Ambitious, Safe, Responsible' with an annex containing five 'Ethical Principles for AI in Defence'.[4]
- On the supranational level, multiple parallel and partly overlapping processes have likewise generated sets of principles:
- GGE LAWS 'affirmed' a set of ten guiding principles in 2018 and added an additional principle in 2019.[5] The 2019 Meeting of High Contracting Parties to the Convention on Certain Conventional Weapons 'endorsed' all eleven guiding principles.[6]

---

[2] US Secretary of Defense, *Memorandum: Artificial Intelligence Ethical Principles for the Department of Defense*, 21 February 2020; US Deputy Secretary of Defense, *Memorandum: Implementing Responsible Artificial Intelligence in the Department of Defense*, 26 May 2021, https://media.defense.gov/2021/may/27/2002730593/-1/-1/0/implementing-responsible-artificial-intelligence-in-the-department-of-defense.pdf..

[3] Ng Eng Hen (Minister for Defence), *Welcome Address* (3rd Singapore Defence Technology Summit, 12 October 2021), https://www.mindef.gov.sg/news-and-events/latest-releases/12oct21_speech; see also Ng Eng Hen (Minister for Defence), *Remarks* (2nd Responsible AI in the Military Domain (REAIM) Summit Ministerial Roundtable, 10 September 2024), https://www.mindef.gov.sg/news-and-events/latest-releases/10sep24_speech2; Singapore, *'Singapore's National Submission on the Topic of Lethal Autonomous Weapons Systems'* (9 May 2024), reproduced in *Lethal Autonomous Weapons Systems: Report of the Secretary-General* (1 July 2024) UN Doc A/79/88, Annex, 101–102.

[4] UK Ministry of Defence, *Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-enabled Capability in Defence* (June 2022), 9–11 and Annex A, https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence.

[5] *Report of the 2018 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems* (23 October 2018) CCW/GGE.1/2018/3, para. 21; *Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems* (25 September 2019) CCW/GGE.1/2019/3, Annex IV.

[6] *Final Report of the 2019 Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects* (13 December 2019) CCW/MSP/2019/9, para. 31 and Annex III ('GGE LAWS').

- In 2021, NATO adopted an AI Strategy, which among other things endorsed six 'Principles of Responsible Use' for AI in Defence.[7]
- In 2023, the US sponsored a Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, which articulates a series of 'measures' that should be implemented in the development, deployment, or use of military AI capabilities.[8] Roughly 60 States have endorsed the Declaration.[9]
- The Netherlands and the Republic of Korea have led two major summits on the Responsible Artificial Intelligence in the Military Domain (REAIM) in 2023 and 2024, which have resulted in two significant public declaration that have likewise attracted the support of about 60 States.[10]

Given the different pedigree and focus of these documents, a few comments may be necessary about the extent to which they can be meaningfully compared.

First, the principles are variously labelled as being 'guiding' or as reflecting 'ethical' or 'responsible' conduct. The precise nomenclature is arguably inconsequential. In all instances, the principles seek to offer some normative but legally non-binding guidance on the development and use of particular military technology. The word 'guiding' does not imply encouragement or a roadmap for developing any technology. The word 'ethical' does not appear to reference any coherent ethical theory; indeed, some of the principles appear to be more prudential than ethical in character.

Second, the US, UK and NATO catalogues of principles are relatively easy to compare to each other because, unsurprisingly, their intent and scope is very similar. Singapore's principles are slightly different in that they ostensibly have a narrower and clearer focus on risk management. At the same time, the latter principles use terminology that mirrors the US, UK and NATO principles.

---

[7] NATO, *'Summary of the NATO Artificial Intelligence Strategy'* (22 October 2021), https://www.nato.int/cps/en/natohq/official_texts_187617.htm; see also Zoe Stanley-Lockman and Edward Hunter Christie, *'An Artificial Intelligence Strategy for NATO'*, *NATO Review* (25 October 2021), https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html.

[8] *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy* (9 November 2023), https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/.

[9] *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy—Endorsing States (as of 27 November 2024)*, https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/.

[10] *REAIM Call to Action* (16 February 2023), https://www.government.nl/documents/publications/2023/02/16/reaim-2023-call-to-action; *REAIM Blueprint for Action*, https://reaim2024.kr/home/reaimeng/board/bbsDetail.do?shareBbsNo=163&shareBbsMngNo=10264&shareMenuId=11613&shareTabMenuId=.

Third, the GGE Guiding Principles do not deal expressly with military AI but are concerned with 'emerging technologies in the area of lethal autonomous weapon systems'. Even that is a generalisation as different principles pertain to 'weapon[s]',[11] 'weapon systems',[12] 'weapon[s], means or method[s] of warfare',[13] 'lethal autonomous weapon systems',[14] 'emerging weapons system[s] in the framework of the CCW',[15] 'emerging technologies in the area of lethal autonomous weapons systems'[16] and — in what can only be described as an affront to the English language — 'weapons systems based on emerging technologies in the area of lethal autonomous weapons systems'.[17] That said, the GGE discussions are often animated by concerns about weapon systems incorporating AI. For example, the challenge of evolving systems and the problem of data bias in the context of machine learning have been persistent issues in the GGE debates. Therefore, it seems fair to say that even though the GGE Guiding Principles are overall not AI-specific, they are pertinent to military AI that may be weaponised.

Fourth, the Political Declaration does not purport to establish 'principles' on the use of military AI. Rather, in its own words, it reflects shared views of the Endorsing States on 'measures [that] should be implemented in the development, deployment, or use of military AI capabilities',[18] thus suggesting a more technical or operational focus. However, the preamble of the Political Declaration and the list of 'measures' have substantial similarities to what are called 'principles' elsewhere. Also, the US State Department's explanatory comments on the Political Declaration confirm that this is not (merely) a technical document but a 'normative framework addressing the use of these [AI] capabilities in the military domain'.[19]

Fifth, the REAIM Blueprint for Action likewise eschews the term 'principle'. Rather, its meatiest section bears the somewhat ambiguous heading 'Implementing responsible AI in the military domain'. That said, in the paragraphs of that section, States 'affirm', 'stress' and 'acknowledge' multiple propositions[20] that substantially overlap with what are articulated as 'principles' in other documents.

---

[11] GGE LAWS (n 6) principle (c).
[12] Ibid, principles (a), (b), (c) and (g).
[13] Ibid, principle (e).
[14] Ibid, principle (a).
[15] Ibid, principle (d).
[16] Ibid, preamble, and principles (h), (i) and (k).
[17] Ibid, principles (c) and (f).
[18] Political Declaration (n 8) preamble.
[19] US State Department—Bureau of Arms Control, Deterrence, and Stability, *'Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy'* (2025), https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/.
[20] REAIM Blueprint for Action (n 10) paras 7–12.

United States	UK	NATO	Singapore	GGE LAWS	Political Declaration	REAIM Blueprint
		lawfulness		IHL 'continues to apply fully'	'compliance with applicable international law'	'accordance with national and international law'
	human-centricity					human-centric
responsible	responsibility	responsibility and accountability	responsible	human-machine interaction	context-informed judgments	appropriate human involvement
				human responsibility, accountability	accountable; oversee	responsible and accountable
equitable	bias and harm mitigation	bias mitigation		risk assessments and mitigation measures	minimise unintended bias and accidents; mitigate risks of failures	reduce risk of malfunctions or unintended consequences, including from data bias
			reliable			
reliable	reliability	reliability			explicit, well-defined uses; safety, security and effectiveness	reliability and trustworthiness
			safe			
			robust			
traceable	understanding	explainability and traceability			transparent and auditable; understand	understand, explain, trace and trust; explainability and traceability
governable		governability			ability to respond	

Table 1: Principles on the ethical or responsible use of military AI

3. The Individual Principles

As Table 1 above illustrates, the sets of principles have important commonalities in terms, concepts and substantive requirements. At the same time, there are a few notable divergences. What follows is a brief comparison of the articulation of the principles, and a general reflection on any linkages with existing law.

3.1 Lawfulness

The principle of lawfulness appears in each of the multinational sets of principles. [21] While the precise language varies slightly, the principle in all instances encapsulates the requirement to **develop and use military capabilities in accordance with applicable law**. NATO Principles and the REAIM Blueprint reference both national and international law, while the GGE Guiding Principles and the Political Declaration are focused on international law. The applicable international law is identified with a different level of specificity across the documents, with the NATO Principles, GGE Guiding Principles and Political Declaration specifically referencing international humanitarian law (IHL), and the NATO Principles also expressly mentioning human rights law (albeit hedged with the phrase 'as applicable').

The US, Singaporean or UK principles do not specifically identify the need to comply with the law as a discrete principle but accept this proposition explicitly or implicitly. For example, the document that the UK principles are attached to, identifies ethical constraints alongside safety and legal considerations, and clearly states that the UK MoD's 'development and use of AI technologies will always be in accordance with the body of applicable UK and international law.[22]

Identifying lawfulness as a standalone principle appears to serve the same purpose. On the one hand, it makes it clear that compliance with the law is a necessary but insufficient condition for describing the use of a technology as responsible. By the same token, it means that the principles do not seek to qualify or displace the law but rather build upon any existing legal requirements.

This, however, raises the more difficult question as to whether the other principles are therefore legally inert by being entirely decoupled from the legal framework. But as the discussion further below seeks to demonstrate, some other principles clearly have substantial legal content. Thus, the better reading of the principle of lawfulness seems to be that it flags compliance with the law as one aspect of responsible use but does not

[21] NATO (n 7) principle A; GGE LAWS (n 6) principle (a); Political Declaration (n 8) preamble; REAIM Blueprint for Action (n 10) para 9(b).
[22] UK MoD (n 4) 6.

preclude the possibility that the other principles might be helpful for understanding the existing law or potentially even progressively developing the law.

3.2 Human-Centricity

The UK principles include the principle of human-centricity, which requires assessing and considering the impact of AI-enabled systems on humans, including the full range of positive and negative effects across the entire system lifecycle.[23] The REAIM Blueprint suggests that 'AI applications should be ethical and human-centric',[24] without offering any explanation as to what that means.

The notion of human-centricity originates from discussions about general AI governance. For example, the Japanese government's 2019 document 'Social Principles of Human-Centric AI' uses human-centricity both as an overarching paradigm for AI governance as well as a specific governance principle.[25] In the latter sense, the principle encapsulates a veritable constellation of ideas, such as the non-infringement of fundamental to human rights guarantees, the use of AI to expand human abilities rather than replace humans, the recognition of human agency and responsibility for the use of AI, and the avoidance of technological divides.[26]

The UK principles appear take a narrower approach by focusing on the assessment of the effects of the system. But even so, the principle extends well beyond the existing legal framework that governs the use of technology by armed forces.

IHL and international human rights law (IHRL) plainly require the assessment of the effects of the use of technology where this may cause harm of the kind that the law seeks to prevent. For example, any use of technology that has the potential to cause injury or death to civilians, would need to be assessed in light of, at the very least, the constant care principle in IHL or the right to life under IHRL.

On the other hand, in giving effect to the principle of human centricity, *JSP 936 Dependable Artificial Intelligence (AI) in Defence* specifies:
All humans (e.g. MOD personnel, civilians, targets of military action etc.) interacting with or affected by the development and/or use of an AI-enabled system must be clearly identified. An assessment must then be made of the impact the AI could have on each

[23] UK MoD (n 4) first principle.
[24] REAIM Blueprint for Action (n 10) para 9(a).
[25] Cabinet Secretariat (Japan), *'Social Principles of Human-Centric AI'* (2019), https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf.
[26] Ibid 7.

stakeholder group to ensure that effects are as positive as possible and justified as outweighing negative effects where these may arise.[27]

Thus, the principle of human-centricity goes much further than the IHL and IHRL requirements by capturing more subtle consequences of the use of AI. This includes, for example, the problem of automation bias (even where that does not lead to unlawful consequences) and the de-skilling of the military workforce.

3.3 Accountability and Responsibility

The notion of accountability and/or responsibility makes an appearance across all the documents considered here. However, it has at least to different connotations, which demand separate attention.

3.3.1 Human Involvement

As used in the documents reviewed, the notions of accountability and/or responsibility relate, first, to **the kind of human involvement required in the use of AI systems**. Thus, under the heading of accountability and/or responsibility, the US and NATO principles contemplate humans exercising 'appropriate levels of judgment and care'[28] in the development and use of AI capabilities, and the UK principles require 'clearly defined means by which human control is exercised throughout their lifecycles'.[29] The GGE Guiding Principles and Political Declaration mention the use of systems during military operations within either 'a responsible chain of human command and control'[30] or 'a responsible human chain of command and control'.[31]

Other references to the desirable human involvement in these and other documents are detached from the notions of accountability and responsibility. For example, another GGE Guiding Principle focuses on 'human-machine interaction' without directly linking it to the notion of accountability or responsibility. [32] Likewise, the Political Declaration refers to human oversight,[33] and the REAIM Blueprint to human involvement, judgment and control, [34] without using the terms accountability or responsibility.

[27] UK Ministry of Defence, *JSP936 Dependable Artificial Intelligence (AI) in Defence—Part 1: Directive* (V1.1, November 2024), para. 52.
[28] US Secretary of Defense (n 2) principle 1; NATO (n 7) principle B.
[29] UK MoD (n 4) second principle.
[30] GGE LAWS (n 6) para (d).
[31] Political Declaration (n 8) preamble.
[32] GGE LAWS (n 6) para (c).
[33] Political Declaration (n 8) measure C.
[34] REAIM Blueprint for Action (n 10) para 9(e).

From an international law perspective, identifying the human involvement requirement presents a challenge. The decade-long debate about autonomous weapon systems shows that while there is broad agreement about the need for human involvement in the lawful use of weapon systems, it has proven difficult to articulate a general statement as to the necessary timing, level and quality of human involvement, and to agree on the terminology to describe such involvement. The problem is that:
IHL does not deal explicitly with the notion of human intervention in [the] operation of weapons. It deals in depth with control over weapons by defining a range of legal obligations the observance of which necessarily requires a high level of control over the application of force, but not specifically with the degree of human intervention that must be employed in applying that control.[35]

With this in mind, the principles have the potential build upon or develop the law by clarifying the human intervention requirement. However, it is doubtful whether any of the documents examined here succeed in doing that. Many of the principles acknowledge that human intervention relates to the entire development, deployment and use lifecycle of a system, thus helpfully dispelling any possible misconception about human intervention being solely or principally a question about a 'real-time trigger-puller'. But it is unclear whether the principle add much beyond this point. References to 'appropriate human involvement', 'appropriate levels of judgment and care' and 'responsible chain of human control' seem to raise as many questions as they answer. The GGE's most recent language, which refers to 'context-specific human control and judgment', aims to bring some of these concepts closer together, but likewise requires further clarification and operationalisation.

The problem may well be that it is impossible to make a statement that would be more granular or practical while at the same time remaining valid for all AI systems. The timing, level and quality of human involvement required for legal compliance appears to depend heavily on the capabilities of the system and the operational environment. From a legal perspective, perhaps the best that could be said is that humans and States must exercise such control and judgment in relation that AI systems as may be necessary for them to comply with their legal obligations.

3.3.2 Consequences of Actions

Accountability and responsibility also refer to **the relationship of the human to the effects or the consequences of the use of an AI capability**. This is expressed in the most straightforward way in the US principles, which indicate that DoD personnel 'remain[] responsible for the development, deployment, and use of AI capabilities',[36] and the Singaporean principles, which stipulate that 'both developers and users are

[35] Tim McFarland, 'Minimum Levels of Human Intervention in Autonomous Attacks', *Journal of Conflict and Security Law* 27, no. 3 (1 September 2022): 387–409, https://doi.org/10.1093/jcsl/krac021.
[36] US Secretary of Defense (n 2) principle 1.

responsible for the outcomes of artificial intelligence systems'.[37] The Political Declaration meanwhile postulates simply that '[m]ilitary use of AI capabilities needs to be accountable'. [38]

Other principles seek to explicate in some way the relationship between the two concepts used, especially by defining responsibility as an aspect of, or means for achieving, accountability. Thus, as per the UK principles, '[h]uman responsibility for AI-enabled systems must be clearly established, ensuring accountability for their outcomes'. [39] Under the NATO principles, 'clear human responsibility shall apply in order to ensure accountability'. [40]

The GGE Guiding Principles and the REAIM Blueprint underscore that accountability and responsibility cannot be transferred to machines.[41] The REAIM Blueprint hedges its bets by using 'responsibility and accountability' in tandem and without attempting the pin down the difference between the concepts.[42] The GGE Guiding Principles seem to use the words accountability and responsibility interchangeably.[43]

None of the documents define what accountability and responsibility mean, which gives rise to some confusion, especially in the GGE Guiding Principles, and makes it more difficult to establish the extent to which they reflect existing international law. With that caveat, a few observations can be made.

Accountability appears to be the broader and looser term. It refers to the possibility of an actor being somehow held to account for their behaviour, which might especially involve being required to explain and justify their behaviour to someone else.[44] Responsibility means that some entity can be blamed for some prohibited or otherwise undesirable behaviour. For the purposes of international law, this relates to the breach of an obligation under international law and the legal consequences that follow from that breach.

Thus, according to the ILC's Articles on State Responsibility, 'every internationally wrongful act of a State entails the international responsibility of that State'. As the Articles make it clear, '[t]here is an internationally wrongful act of a State when conduct consisting of an action or omission: (a) is attributable to the State under international law; and (b) constitutes a breach of an international obligation of the State.'

---

[37] Singapore (n 3) principle a.
[38] Political Declaration (n 8) preamble.
[39] UK MoD (n 4) second principle.
[40] NATO (n 7) principle B.
[41] GGE LAWS (n 6) para (b); REAIM Blueprint for Action (n 10) para 9(c).
[42] REAIM Blueprint for Action (n 10) para 9(c).
[43] GGE LAWS (n 6) para (b), (c) and (d).
[44] Jan Klabbers, *International Law* (Cambridge University Press, 2024), p. 136.

There is no equally authoritative and concise statement of acts that entail individual responsibility under international law. Extrapolating from the jurisprudence of the ICTY on war crimes, one can perhaps say (with some circularity) that a person can be held criminally responsible under international law where the conduct of a person constitutes an infringement of an applicable rule of international law, the violation is serious, and the violation entails, under international law, the individual criminal responsibility of the person breaching the rule.[45] Individual responsibility as a matter of domestic law, insofar as it is contemplated by the various principles, could entail punishment under domestic criminal law or military discipline law, or some form of administrative action.

In this second sense, the principle of accountability and responsibility has a significant link to existing law. For one, it serves as a reminder of the consequences of breaches of international law. Also, international law may specifically require individuals to be held accountable. Notably, States must 'repress' grave breaches of the Geneva Conventions and Additional Protocol I.[46] This is generally taken entail an obligation to search for and try persons accused of having committed or having ordered the commission of such breathes.[47] In other words, this is an obligation to establish individual criminal responsibility. At the same time, States must 'supress' all other acts contrary to the provisions Geneva Conventions and Additional Protocol I. This generally refers to broader range of measures, such as administrative inquiries.[48] Thus it contemplates accountability in a broader sense than individual criminal responsibility.

3.4 Bias and Harm Mitigation

The US, UK, Singapore and NATO principles, and the Political Declaration, expressly require **the taking of proactive measures to minimise, reduce and/or mitigate unintended bias** in AI capabilities or applications.[49] In this context, the NATO principles refer expressly to bias in 'data sets'[50] and the Blueprint references 'data, algorithmic and other biases'.[51]
Beyond bias, the principles call for **addressing other undesirable consequences**:

- the UK principles refer to mitigating the risk of unexpected or unintended 'harms';[52]

[45] Compare *Tadić Jurisdiction Decision* (ICTY, Case No IT-94-1, 2 October 1995), [94].
[46] GC I, art 49; GC II, art 50; GC III, art 129; GC IV, art 146 and AP I, art 85(1).
[47] See, eg, International Committee of the Red Cross (ICRC), *Guidelines on Investigating Violations of International Humanitarian Law: Law, Policy, and Good Practice* (Geneva: ICRC, 2019), https://www.icrc.org/sites/default/files/document/file_list/guidelines_on_investigating_violations_of_ihl_final.pdf, para 16.
[48] Ibid.
[49] US Secretary of Defense (n 2) principle 2; UK MoD (n 4) fourth principle; Singapore (n 3) principle b; NATO (n 7) principle F; Political Declaration (n 8) measure D.
[50] NATO (n 7) principle F.
[51] REAIM Blueprint for Action (n 10) para 9(d).
[52] UK MoD (n 4) fourth principle.

- the Singaporean principles require addressing 'the risk of errors or inaccuracies in an artificial intelligence system's output';[53]
- the Political Declaration requires minimising 'accidents', mitigating 'risks of failures', and detecting and avoiding 'unintended consequences';[54]
- the REAIM Blueprint refers to reducing the risk of 'malfunctions or unintended consequences'.[55]

Interestingly, in terms of the taxonomy, Singapore addresses these issues as an aspect of reliability, whereas the other principles treat this as a discrete principle.

IHL's general regulatory approach is not to require harm minimisation or mitigation as such but to prohibit or require certain conduct, including through the setting of certain thresholds. Thus, in the conduct of hostilities, IHL requires directing attacks only against lawful objectives (principle of distinction) and refraining from attacks which may be expected to cause collateral damage that is excessive in relation to the concrete and direct military advantage anticipated (rule of proportionality). Importantly, proportionality under IHL is a specific rule setting limits on collateral damage. It is not an overarching principle of the law[56] that would, for example, require minimising all kinds of undesirable consequences of military action.

That having been said, IHL comes reasonably close to containing a specific requirement of civilian harm reduction in attack. Most obviously, under Article 57(2)(a)(ii) of Additional Protocol I, those who plan or decide upon an attack must, *inter alia*, take all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event to minimising, incidental loss of civilian life, injury to civilians and damage to civilian objects. Thus, there is a distinct duty to reduce incidental harm to civilians though a choice of means of methods of warfare, provided that the choice is practicable or practically possible, considering all circumstances ruling at the time, including humanitarian and military considerations.[57]

The principle of harm mitigation appears be far more comprehensive than corresponding obligations that exist under existing law. In particular, it captures a far broader range of undesirable consequences than the direct civilian harm that is a major focus of the conduct of hostilities rules under IHL. Also, it extends beyond the context of attacks, which are the focus on Article 57(2).

---

[53] Singapore (n 3) principle b.
[54] Political Declaration (n 8) measure J.
[55] REAIM Blueprint for Action (n 10) para 9(d).
[56] See Jeroen van den Boogaard, *Proportionality in International Humanitarian Law: Refocusing the Balance in Practice* (Cambridge University Press, 2023).
[57] For this definition of 'feasibility', see, eg, *Protocol on Prohibitions or Restrictions on the Use of Incendiary Weapons (Protocol III)* (10 October 1980), art. 1(5)..

3.5 Explainability and Traceability

Explainability and traceability are used across the different sets of principles with some degree of overlap. When considered together, they appear to have two main aspects. First, relevant individuals must have an appropriate understanding of the capabilities. What precisely this understanding relates to has been expressed with a different level of granularity in the different sets of principles. Thus, the subject matter variously includes:

- 'the technology, development processes, and operational methods',[58]
- 'AI-enabled systems, and their outputs',[59]
- 'the capabilities and limitations of those systems',[60] and
- 'the outputs produced by AI capabilities'.[61]

Second, explainability and traceability also reflect transparency and auditability with respect to 'methodologies, data sources, and design procedures and documentation',[62] or through the use of 'review methodologies, sources, and procedures', which includes 'verification, assessment and validation mechanisms'.[63]

IHL does not require individuals to have any specific level of understanding of the systems that they utilise, particularly as regards the technical nuances of their operation. For example, there is no legal requirement for a combatant to have an in-depth understanding of projectile motion physics or wound ballistics, even though some knowledge of these matters would no doubt be operationally beneficial. That said, the ability to understand the effects of systems, including their capabilities and limitations, appears to be an essential precondition for compliance with certain rules of the law.

The principle of precaution provides the broadest example. Article 57(1) of Additional Protocol I stipulates that '[i]n the conduct of military operations, constant care shall be taken to spare the civilian population, civilians and civilian objects'. It is difficult to see how constant care could be taken in circumstances where military personnel do not fully grasp the ways in which the operation of a system in the context of operations could have an adverse impact on the civilian population. Constant care requires such impact to be actively considered and minimised, which is plainly dependant on an understanding of what the system can and cannot do.
More specifically, under Article 57(2)(a)(ii) of Additional Protocol I, those who plan or decide upon an attack must, *inter alia*, take all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event to minimizing,

[58] US Secretary of Defense (n 2) principle 3.
[59] UK MoD (n 4) third principle.
[60] Political Declaration (n 8) measure G.
[61] REAIM Blueprint for Action (n 10) para 9(f).
[62] Political Declaration (n 8) measure F; similarly US Secretary of Defense (n 2) principle 3.
[63] NATO (n 7) principle C.

incidental loss of civilian life, injury to civilians and damage to civilian objects. To make an informed decision about the suitable means and methods of warfare, the operator must consider the comparative advantages of different weapon systems in minimising collateral harm. Again, this clearly presupposes an understanding of the effects of weapon systems.

3.6 Reliability

Reliability has at least two dimensions. On the one hand, as suggested in the US and NATO Principles, and the Political Declaration, it means that **AI capabilities have explicit and well-defined use cases, and that they are designed and engineered to fulfill those intended functions**.[64]

On the other hand, as articulated in almost all the documents examined—albeit with varying focus and granularity—reliability includes **safety, security, robustness and effectiveness of the AI capabilities**.[65] This ought to be ensured through appropriate safeguards, such as rigorous testing and assurance within their well-defined uses and across their entire life-cycles, and monitoring to detect performance degradation.

The Singaporean principles take a unique approach. As already mentioned, they treat bias and risk mitigation as an aspect of reliability. Furthermore, Singapore has identified robustness and safety as discrete principles, while the other sets of principles refer to these concepts in general terms under reliability. The Singaporean approach is therefore quite instructive. To ensure robustness, 'the risks from the exploitation of artificial intelligence by malicious actors must be addressed' and AI capabilities 'should be designed with cyber and adversarial artificial intelligence threats in mind'.[66] In terms of safety, Singapore's principles direct 'focus on the risk of [AI] failure in safety-critical contexts'; also AI capabilities 'should be safe to use, not only in terms of the deployed platforms, but also for the surrounding assets and personnel'. [67]

IHL does not expressly require weapons, means of warfare or other capabilities to be reliable. However, as articulated in the previous section in relation to the need to understand the effects of the system, compliance with certain rules of the law requires operators to anticipate and control the effects of weapon systems. Where the system is unreliable, it is difficult to see how the operator could comply with the principle of precaution, including both the constant care obligation and the requirement to take specific precautionary measures in attack.

---

[64] NATO (n 7) principle D; Political Declaration (n 8) measure H.
[65] US Secretary of Defense (n 2) principle 4; UK MoD (n 4) principle 5; NATO (n 7) principle D; Political Declaration (n 8) measure I; REAIM Blueprint for Action (n 10) para 9(d).
[66] Singapore (n 3) principle (c).
[67] Ibid, principle (d).

At the same time, the principle of reliability also incorporates safety vis-à-vis the operating personnel. This is not a matter addressed by IHL. But, to a significant extent, human rights law captures this issue. Courts have indicated that inadequacy of equipment that contributes to the death of members of armed forces may amount to an interference with, and the breach of, their right to life.[68]

3.7 Governability

Governability appears as a discrete principle in the US and NATO principles, as well as the Political Declaration. In each of these, it has two complementary aspects: [69]

- the ability to detect and avoid unintended consequences; and
- the ability to respond to such unintended behaviour, in particular by disengaging or deactivating the system.

The US and NATO principles connect these aspects of governability to the capacity of the system to be used fulfill its intended functions, whereas the Political Declaration presents them as safeguards to mitigate risks of failures.

Governability does not directly correlate to a specific rule of LOAC. It could be seen as deriving from the general obligation to respect and ensure respect for LOAC,[70] but the principle extends further, covering unintended consequences that are not necessarily breaches of LOAC. For example, an AI system that tends to trigger friendly fire incidents would not be inconsistent with LOAC, as no LOAC rules protects against friendly fire, but such a system would nevertheless by problematic in terms of governability. Governability could also be seen as a facet of the obligation to take constant care in military operations to spare the civilian population, civilians and civilian objects.[71]

When it comes to weapons, means and methods of warfare, governability could be grounded in some further rules. A weapon falls foul of the prohibition of inherently indiscriminate weapons if it strikes military objectives and civilians or civilian objects without distinction as a consequence of the inability to limit its effects as required by law.[72] A weapon system characterised by unexpected behaviour without the possibility of deactivation could be seen as having effects that cannot be limited. Also, such a system, when it operates over a longer period of time, makes it difficult to comply with the obligation to cancel or suspend an attack when inconsistencies with the principle of distinction or the rule of proportionality become apparent.[73]

[68] See *Smith v Ministry of Defence* [2013] UKSC 41.
[69] US Secretary of Defense (n 2) principle 5; NATO (n 7) principle E; Political Declaration (n 8) measure J.
[70] See GC I–IV, common art 1; AP I, art 1(1).
[71] AP I, art 57(1).
[72] AP I, art 51(4)(c).
[73] AP I, art 57(2)(b).

4. Conclusion

The principles on responsible use of military AI identified in various national and international governance instruments have significant similarities. The taxonomy of principles is broadly similar, and the normative content of the principles has notable overlaps across the various instruments. However, the correlation is not perfect. For example, while the principle of reliability is widely recognised, the principles are differently articulated.

The first conclusion is, therefore, that while it may be appropriate to include the principles on the responsible use of military AI in new regulatory instruments (such as the instrument developed by the GGE), this should be done with deliberation and precision. In particular, simply referring to the principles by their 'labels' as a list of '-ilities' may not be the most appropriate approach. If the principles are included, their meaning should be unpacked in the appropriate parts of the instruments, such that there is no confusion as to what they mean.

A further issue is that the principles can misconstrue and oversimplify the law. It is attractive and easy to say that humans need to remain responsible and accountable for their use of AI systems. But the way in which criminal law, for example, applies to misconduct in the context of an armed conflict can be quite complicated. There are complex tests about individual criminal responsibility and command responsibility, to say nothing of the procedures for holding people responsible. Therefore, it is not simply a matter of agreeing that humans shall be responsible and accountable. The question is about the circumstances and the way in which that responsibility and accountability materialise.

Moreover, the principles can conflate an actual legal obligation, the necessary means for complying with a legal obligation, and what is prudent and practical in certain circumstances. For example, while traceability of AI processes might not be strictly required in order to comply with any rule of international law, it might be very helpful for ensuring such compliance and establishing accountability in case of mishaps.

The second conclusion is, therefore, that any inclusion of the principles on the responsible use of military AI should articulate their relationship to the existing legal framework. In some instances, application of the principles may be deemed conducive to compliance with the existing law, whereas in other cases they may significantly broaden the scope of existing obligations.

Finally, it is interesting to note that none of the documents discussed here identify predictability as a discrete principle or even an aspect of one of the other principles. At the same time, predictability has been frequently mentioned in the GGE discussions. For one, it has been proposed by the ICRC as standard for distinguishing between acceptable and unacceptable autonomous weapon systems. It has also been incorporated into Draft Protocol VI and proposed for including in the GGE's current rolling text.[74]

The third conclusion is that, in the absence of guidance in existing principles on the responsible use of military AI, there needs to occur a substantive discussion on the meaning of this standard and the ways of assuring compliance with it. Existing principles probably see predictability as a dimension of reliability, but the precise relationship between these two concepts would benefit from further clarification.

[74] *CCW Draft Protocol VI* CCW/GGE.1/2023/WP.6 (10 May 2023), art. 4(1); *CCW GGE LAWS Rolling Text* (6 March 2025), section III(6).

About the Author

Prof. Dr. Rain Liivoja

Rain Liivoja is a Professor and Deputy Dean (Research) at the University of Queensland Law School, where he leads the Law and the Future of War research group. Rain is also a Senior Fellow with the Lieber Institute for Law and Land Warfare at the United States Military Academy at West Point. He holds the title of Adjunct Professor of International Law at the University of Helsinki, where he is affiliated with the Erik Castrén Institute of International Law and Human Rights.

HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
@hcssnl

The Hague Centre for Strategic Studies