Sweijs, Tim; Romansky, Sofia

Working Paper

International norms development and AI in the military domain

CIGI Papers, No. 300

Provided in Cooperation with:
Centre for International Governance Innovation (CIGI), Waterloo, Ontario

Suggested Citation: Sweijs, Tim; Romansky, Sofia (2024) : International norms development and AI in the military domain, CIGI Papers, No. 300, Centre for International Governance Innovation (CIGI), Waterloo, ON, Canada

This Version is available at:
https://hdl.handle.net/10419/303159

Centre for International Governance Innovation

The Hague Centre for Strategic Studies

CIGI Papers No. 300 — September 2024

International Norms Development and AI in the Military Domain

Tim Sweijs and Sofia Romansky

Centre for International Governance Innovation

The Hague Centre for Strategic Studies

CIGI Papers No. 300 — September 2024

International Norms Development and AI in the Military Domain

Tim Sweijs and Sofia Romansky

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

About the HCSS

The Hague Centre for Strategic Studies is a knowledge institute that conducts independent research on geopolitical, defence and security issues to governments, international institutions and businesses. Our research is characterized by a data-driven, multidisciplinary approach, specialist expertise and a strategic orientation. We combine broad, conceptual knowledge with qualitative and quantitative methods and present our findings in the form of recommendations, strategic explorations and scenario analyses. Our goal is to offer fact-based analysis of the challenges that our societies face in order to inform public discourse, public and private strategic decision making, and contribute to international and national security in accordance with liberal democratic values.

Credits

Managing Director and General Counsel Aaron Shull
Director, Program Management Dianna English
Program Manager and Research Associate Kailee Hilt
Publications Editor Susan Bubak
Publications Editor Christine Robertson
Graphic Designer Sami Chouhdary

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Table of Contents

About the Authors

Tim Sweijs is the director of research at The Hague Centre for Strategic Studies (HCSS) and a senior research fellow at the War Studies Research Centre of the Netherlands Defence Academy. He is the scientific advisor to the Secretariat of the Global Commission on Responsible Artificial Intelligence in the Military Domain, an initiative of the Dutch Ministry of Foreign Affairs.

For close to two decades, Tim has advised international organizations, governments and defence departments across the globe. He has provided expert testimony to the United Nations Security Council, the European Parliament and the Dutch Parliament, as well as to the North Atlantic Treaty Organization's Parliamentary Assembly.

At HCSS, Tim is responsible for the overall research portfolio of the entire institute. He is also a research affiliate in the Center for International Strategy, Technology and Policy at the Sam Nunn School of International Affairs at the Georgia Institute of Technology in the United States. He holds degrees in war studies (Ph.D., M.A.), international relations (M.Sc.) and philosophy (B.A.) from King's College London and the University of Amsterdam.

Sofia Romansky is a junior strategic analyst at HCSS. Her primary research interests concern the impact of artificial intelligence (AI) on social and military domains, issues around narratives and disinformation in online spheres, and Russia's invasion of Ukraine.

On the topic of AI, Sofia has researched the hardware and software components that enable the application of AI to autonomous systems and the impact of post-2018 generative AI on social stability. Within these projects, her focus has been primarily on computer vision and multimodal generative models.

Sofia holds a bachelor's degree (cum laude) in politics, psychology, law and economics with a specialization in politics and a minor in communication science from the University of Amsterdam. She also holds a master's degree (cum laude) in international relations and diplomacy from Leiden University and Clingendael Institute.

Acronyms and Abbreviations

AI	artificial intelligence
AUDA-NEPAD	African Union Development Agency-New Partnership for Africa's Development
GGE on LAWS	Group of Governmental Experts on Lethal Autonomous Weapons Systems
NATO	North Atlantic Treaty Organization
OECD	Organisation for Economic Co-operation and Development
OODA	observe-orient-decide-act
R&D	research and development
REAIM	Responsible Artificial Intelligence in the Military Domain
UNGA	United Nations General Assembly
UNODA	United Nations Office for Disarmament Affairs

Executive Summary

The integration of artificial intelligence (AI) into the military domain is rapidly becoming a reality. This development has the potential to not only affect the character of war but also recalibrate the strategic calculations of political actors in peace time. The breadth and diversity of implications brought forth by military applications of AI has prompted governments and international organizations to formulate norms that would steer the development and use of AI toward adherence with fundamental legal and ethical principles. To wit, no less than seven international initiatives exist dedicated to governing military AI, while various others aim to steer the development and use of AI more broadly. These efforts at norm development and norm setting confront a number of challenges related to the nature of AI applications and to pressures associated with interstate competition. Yet these initiatives have also outlined the key principles that will continue to guide the evolving governance of AI in the military domain. This paper distinguishes seven focus areas found in international governance initiatives and identifies their strengths and weaknesses, as well as overlaps and gaps in the emerging normative landscape. To remedy the identified weaknesses and gaps, this paper argues that it is important to lay the normative foundations for future norms developments now by going beyond conceptual issues while delving into technical and operational specifics. Simultaneously, it is essential to start creating an institutionalized regime of norms, rules and regulations to guide state behaviour, focusing on the entire production-proliferation-deployment-employment chain. These elements will contribute to a robust governance framework for AI in the military domain.

Introduction

Advances in AI have started to percolate into the military domain, at first gradually and at present more rapidly. Military AI applications not only affect the conduct of war in conflict theatres around the world, but they also reshape the dynamics of security competition in peace time. Given the breadth of the current and potential effects of military AI applications, some governments and international organizations have started formulating norms to steer the responsible development and use of these tools in alignment with fundamental legal and ethical principles (Canca 2023, 59; Anand and Deng 2023, 20). At the same time, considerations concerning potential tactical and strategic advantages to be derived from military AI also inform the positions of actors in this realm. In addressing the challenges posed by military AI, some argue that new norms are not necessary as the international community should solely apply the rules and regulations already enshrined in international humanitarian and human rights law. A focus on developing new norms would only dilute the discussion, drawing attention and resources away from the effort to guide responsible uses of AI (Maas and Villalobos 2023, 8; Garcia 2023, 204). Others contend that while these rules and regulations certainly apply, the emergence of military AI necessitates normative elucidation of how these rules and regulations apply. Ensuring the responsible use of AI requires regulatory efforts that transcend the battlefield and extend to the entire production-proliferation-deployment-employment chain (van Hooft, Boswinkel and Sweijs 2022; Scharre 2023). Norms to control, curtail and delineate military AI applications would not affect the core tenets of international law. Rather, they would build on them, as well as expand and refine the scope of existing normative efforts.

Historically, once new technologies "come online," political entities have formulated new guidelines for their use.[1] Political actors voluntarily develop international norms for a number of reasons, including but not limited to *identity-based and moral values* (to prevent human suffering and ensure technologies are used in accordance with basic precepts of international law), *strategic* interests (to maintain a perceived advantageous position of power and to ensure that these technologies do not undermine overall security), and *stability* interests (to ensure that these technologies do not undermine overall [strategic] stability and thereby threaten their security). Ultimately, for norms to have a tangible effect on the behaviour of international actors by defining collective expectations, norms should be adoptable, verifiable and enforceable (Finnemore and Sikkink 1998, 891). Although norms are less binding than international law, they inform the "rules of the game" in the international arena. Yet the formulation and implementation of norms in relation to military AI faces several challenges intrinsic to the nature of AI as an all-purpose technology and the dynamics of interstate competition. As such, initiatives attempting to govern military AI should address the diverse nature of military AI applications while being sufficiently flexible to account for the constant improvement of AI systems in a competitive world.

To wit, no less than seven international initiatives exist dedicated to governing military AI, while various others aim to steer the development and use of AI more broadly. Unsurprisingly, international initiatives often acknowledge that, in accordance with existing international law, military AI should be used in a way that mitigates harm (Boulanin and Lewis 2023, 6; Vestner 2022; D. A. Lewis 2022). Beyond this, however, there is little agreement that could meaningfully guide state behaviour.[2] Other than outlining 11 guiding principles in 2019, the deliberations of the Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE on LAWS) have "proven to be slow and difficult due to the lack of consensus on agenda items" (Schmitt 2022, 306) and the diverging views of key players such as China, Russia and the United States (Bode et al. 2023, 9). Concurrently, the call to action of the Summit on Responsible Artificial Intelligence in the Military Domain (REAIM) and the United States' Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy have been primarily criticized for failing to define what "responsible AI" entails (Nadibaidze 2023). Finally, the representativeness and thereby relevance of initiatives led by, for example, the Organisation for Economic Co-operation and Development (OECD) has been questioned due to its limited membership and domination by the Group of Seven (von Ingersleben-Seip 2023, 797).

This paper takes stock of international norm development around military AI. Following this introduction, the next section offers a brief explainer of the use and utility of international norms and the conditions for successful norm formulation and propagation. The following section titled "Challenges to International Norm Development for AI in the Military Domain" identifies six challenges for norm development specifically with respect to military AI. The next section, "International Norms on AI in the Military Domain: Taking Stock of a Crowded Landscape," maps 13 normative initiatives. The following section titled "An Assessment of Strengths and Weaknesses in International Norms for AI in the Military Domain" distinguishes seven focus areas and identifies strengths and weaknesses as well as overlaps and gaps in the emerging normative landscape. The final section concludes by outlining a preliminary agenda for norms development for AI in the military domain. It argues that it is important to lay the normative foundations for future norms developments now by going beyond conceptual issues while delving into technical and operational specifics, to start creating an institutionalized regime of norms, rules and regulations to guide state behaviour, that should focus on the entire production-proliferation-deployment-employment chain.

1 Efforts include constraints on the use of the crossbow in eleventh-century Europe and handguns under the seventeenth-, eighteenth- and nineteenth-century shogunate in Japan; submarines, balloon-delivered projectiles and expanding bullets (in The Hague Conventions of 1899 and 1907); and poisonous gasses and chemical weapons (from very early attempts dating back to antiquity to the Biological Weapons Convention in 1972 and the Chemical Weapons Convention in 1993); on to land mines (1997) and cluster munitions (1868, 2008); on to nuclear arms control treaties between the United States and Soviet Union to curtail the risks associated with nuclear weapons by placing numerical caps and establishing means of national verification; and broader treaties such as the Missile Technology Control Regime in 1987 and the Wassenaar Arrangement in 1996 to control the proliferation of dual-use technologies. For an overview, see the Annex; Scharre and Lamberth (2022).

2 This observation is not only applicable to AI in the military domain, but also attempts to govern AI more broadly. See Munn (2023, 870).

The Use and Utility of International Norms

International Norms: Definition, Function and Pathways

In the international arena, norms are one tool that states can use to establish some degree of predictability in an otherwise highly uncertain setting. International norms emerge and evolve through repeated interstate interactions where mutual adherence is dependent on trust and common understanding. Norms are not legally binding but can form the foundation for normative regimes: agreements that go beyond temporary arrangements that are sensitive to current state interests (Krasner 1982). In this context, norms are part of a process of reiterating and shaping values and interests, rather than a stand-alone commodity (Klimburg and Almeida 2019). This observation also reflects the difference between social and legal norms, where social norms reflect "intersubjective understandings of 'appropriateness,'" which become legal norms only upon codification in laws (Bode 2023, 42). Social norms can be seen as a dynamic and flexible counterpart to codified law that can help standardize behaviour in areas not yet captured by regulation. Simultaneously, the level at which norms are propagated also matters. Norms promoted by governments to designate state-level behaviour constitute "big N" norms, whereas the standards and protocols set by non-state actors contribute to "small n" norms, which may reflect the values and interests of states and constrain the behaviour of actors (Faesen et al. 2021, 13–15).

How and why norms emerge is a field of study on its own. Particularly insightful and foundational has been the work of Martha Finnemore, who described the norm life cycle in terms of *emergence, cascade and internalization* (Finnemore and Sikkink 1998, 895). During norm emergence, "norm entrepreneurs" seek to anchor the thinking on a norm by proposing specific framing and formulation. As the initial norm is propagated, discussed and accepted, the norm evolves to reflect more than just the norm entrepreneur's framing. If successful, a cascade occurs once a critical number of key actors accept the norm, increasing the appeal for others to follow. In turn, norm internalization takes place as norms become entrenched in values, interests and behaviour. The rate at which norms mature through this cycle directly depends on the context of a norm, the identity to which it appeals and applies, as well as the corresponding behaviour and expectations. This overall process is non-linear, emerging from multi-level interactions between different actors (Winston 2023).[3]

Yet the strength of social norms — their voluntary and flexible nature — also constitutes their weakness, even once codification into legal norms has occurred. Some states may engage in "norm signalling," the performative adherence to norms to gain reputational benefits or access to certain discussions, without adjusted behaviour (Dixon 2017). Alternatively, states may deliberately interpret or attempt to shape social and legal norms in the broadest way possible to justify marginal behaviour (Dixon 2017; Farrell 2005). States might also outright violate a norm while spinning the narrative to present their behaviour as being aligned with the corresponding values (S. Cohen 2001). Norm-breaking behaviour may attract condemnation, but often international actors have little tools to sanction behaviour. Various factors impact the extent to which a particular norm is ultimately embraced by the international community, building on strategic and stability interests on the one hand and values on the other. Specifically, three conditions favour the propagation and institutionalization of norms: adoptability, verifiability and enforceability.

Conditions Favouring Norm Propagation

Adoptability

Adoptability can be straightforwardly defined as the extent to which norms are amenable to agreement and approval by key actors based on their interests and values. In the case of military AI, hard-nosed assessments of the potential military advantages stemming from the use of particular AI technologies will be weighed against the extent to which they align with key moral values. This is an inherently social and thereby *strategic* process. As such, norm adoption is influenced by which other actors adopt norms and whether they are

3 For example, the approach used by the Global Commission on the Stability of Cyberspace involved allowing "states and other stakeholders to embrace some norms while rejecting or abstaining from others" to clarify areas of consensus and disagreement, and to foster the embracing of specific norms (Global Commission on the Stability of Cyberspace 2019, 23).

perceived to be partners, allies or potential foes. Additionally, these actors' influence may be rooted in their material and ideational power base, as "norms held by powerful actors simply have many more opportunities to reproduce through the greater number of opportunities…to…persuade others of the rightness of their views" (Florini 1996). Other states may follow the example of norms adopted by powerful states, regardless of the active promotion or content thereof, based on a simple assessment of "the benefits or costs implied in the rule-following action" (Argomaniz 2010, 120). Concurrently, key national values inform decisions about which norms are embraced or rejected (Wendt 1994, 385). For example, a liberal-democratic identity carries with it the prerequisite of adopting particular human rights norms (Gurowitz 2006). In reality, though, governments do not always live up to the values that come with their identity. Neither interests nor values are objectively given, and choices may also result from less rational decision-making processes (Sugden 1989). Appreciation of interests and values is assembled in a complex process involving different stakeholders (Gould, Arentze and Hoijtink 2024).

Verifiability

Verifiability refers to whether the compliance of states with norms can be established, either through formal (for example, institutionalized inspection regimes) or less formal (for example, monitoring by non-state actors) means (Scharre and Lamberth 2022). A norm may be non-verifiable not only because it is defined in terms of conditions or actions that are difficult or impossible to observe, but also because verification requires checking conditions or actions which, albeit technically possible, would require institutional facilitation (Dastani, Torroni and Yorke-Smith 2018).

Regarding military AI applications, secrecy can be an obstacle in ensuring verification of norm compliance. Initiatives in other realms suggest that such obstacles need not be insurmountable and can be addressed using combinations of technological and social confidence-building measures.[4] Verification can also be pursued by other means, including intelligence based on, for example, open-source reporting (Scharre and Lamberth 2022).

Generally, given the software-hardware nexus essential for AI applications, compliance could potentially be verified in one of two ways: either through assessing the technical characteristics of AI systems, or through monitoring their actual case-specific use in the military domain. This process requires concurrent technical, legal and military expertise to conduct legal reviews; creating software protocols; establishing monitoring and inspection mechanisms; as well as implementing confidence-building measures to improve transparency and facilitate legal verification (Goussac et al. 2023).

Enforceability

Enforceability refers to the extent to which norm compliance can be effectuated, usually in tandem with verification of norm adherence or deviation. This way, verification validates an "escalation" of enforcement beyond goodwill (Faesen et al. 2020). Regarding AI in the military domain, enforcement strategies can target the entire life cycle from production and proliferation to deployment and employment. This process may involve controlling the input necessary to create AI applications, from software (algorithms) and hardware (systems) to wetware (people), or countering specific applications in the battlefield through technical "constraining parameters" or "boundary conditions" hardcoded into systems (Sastry et al. 2024, 55; van Hooft, Boswinkel and Sweijs 2022, 81). *Ex post*, it can include public condemnation when actors are in contravention, extend to criminal or state responsibility, and trigger demands for reparation when acts are deemed unlawful (Zyberi 2018; Sassòli 2002; Wolfrum 1987). Other methods of enforcement include bans and moratoria, non-proliferation regimes, export control lists, licensing regimes, tracking of and/or registering key resources, and confidence-building measures to foster mutual trust and the effectiveness of the enforcement regime, even though in the realm of AI, it will likely be difficult to "negotiate such an intense level of oversight" (Maas and Villalobos 2023, 31).[5]

4 The Verification Research Training and Information Centre initiative between Norway and the United Kingdom, for instance, to verify the dismantlement of a mock-up nuclear weapon, is a case in point. See Persbo (2010).

5 For an excellent example of lessons to be learned from confidence building, see Cervasio, Wheeler and McClafferty (2024); Maas and Villalobos (2023); Drexel and Depp (2023).

Challenges to International Norm Development for AI in the Military Domain

The creation of international norms for military AI faces several challenges deriving from the intrinsic characteristics of AI applications and the competitive dynamics that have recently surged in the security and economic realm (R. S. Cohen et al. 2023; Mazarr 2022). Six principal challenges stand out:

- the breadth of AI as an all-purpose technology;
- the difficulty of controlling the inputs that go into AI applications;
- the diversity of actors involved throughout the AI life cycle;
- the challenge of identifying military uses of AI;
- the perception of AI as crucial to attaining a competitive advantage in interstate competition; and
- the AI power paradox, where the rate of technological development outpaces the rate of policy formulation and adoption (see Table 1). (Bremmer and Suleyman 2023)

Norm development around military AI is shaped by these challenges. Therefore, understanding these challenges is necessary to chart the agenda for future norm development.

Table 1: Summary of the Six Challenges Faced by Norm Development for Military AI

Challenge	Description
AI as an all-purpose technology	The concept of AI encompasses an incredibly wide range of multi-purpose technologies and applications, with no universal definition. This necessitates that any normative discussion around AI must start by deciding on a definition that reflects the goals of the initiative and the respective cases where norms would, should and can apply.
The variety of inputs that go into AI applications	AI as a category of technologies is dependent on both software and hardware. As such, any potential controls should account for the interactions and limitations of inputs such as data, computing power and human talent.
The diversity of actors involved in the AI chain	A plethora of actors with different interests and values are involved in the production-proliferation-deployment-employment chain for AI in the military domain. Frameworks need to account for the roles and responsibilities played by different actors that are subject to various controls and guidelines.
The difficulties of verifying the use of AI in the military domain	It is often difficult, if not nearly impossible, to determine whether and how AI has been used in a military context as AI does not necessarily alter the physical characteristics of systems, and systems themselves can switch between AI and non-AI-enabled modes.
The role of AI in interstate strategic competition	Key state actors are reluctant to commit to governance initiatives around AI, lest they lose perceived competitive advantages and because of issue linkage. These considerations make it difficult to arrive at shared positions beyond lowest-common-denominator agreements.
The AI power paradox	The current rate of AI development outpaces the rate at which policies can be formulated and adopted. While AI has the potential to alter the status quo in many areas of life, the urgency of coming up with responses is counterbalanced by the need to create thorough and well-evaluated regulation.

Source: Authors.

Challenge 1: AI as an All-Purpose Technology

First, the concept of AI encompasses a wide range of multipurpose technologies and applications. Most broadly, AI can be understood as a system that can carry out tasks at a level comparable with, or normally dependent on, human intelligence with varying degrees of autonomy and adaptability (Sheikh, Prins and Schrijvers 2023, 16; Russell, Perset and Grobelnik 2023). However, no universally agreed-upon definition of AI exists (Russell and Norvig 2016). Some argue that there is still insufficient understanding of what constitutes human intelligence and, consequently, to what extent machines are successful in its imitation (Sheikh, Prins and Schrijvers 2023, 16). Others argue that as AI technologies evolve, so does the perception of what constitutes AI. For example, the advances made by generative AI models since 2018 have forced people to reconsider the perceived boundaries of AI systems. Still, all modern AI remains "narrow"; for now, artificial general intelligence, which could purportedly execute all forms of human behaviour across all domains, remains to be realized.[6] In the military domain alone, AI can be used for a variety of purposes across the entire observe-orient-decide-act (OODA) loop: from the automated analysis of images and decision-making support in the generation of courses of action, to deployment in the field as part of autonomous weapons systems (Meerveld et al. 2023, 14). Consequently, norm formulation for military AI should start by deciding on a definition of AI that reflects the goals of the normative initiative and the situations where respective norms *would, should* and *can* apply, rather than trying to find a perfect, all-encompassing definition. This first step is crucial for ensuring that resulting frameworks are relevant and effective.

Challenge 2: The Variety of Inputs that Go into AI Applications

Second, AI applications are dependent on an assortment of inputs: data to train AI models, computing power to process the data, human talent to develop algorithms and institutions to guide these interactions following discreet value sets.[7] AI models, training data and resulting algorithms are fundamental to AI applications and can be shared digitally with relative ease. Generative AI applications, such as OpenAI's ChatGPT, Google's T5 and Facebook's LLaMA, were trained on data scraped from publicly available online information (Schreiner 2023; Schaul, Chen and Tiku 2023). Other models, such as iterations of the YOLO (You Only Look Once) computer vision model, are downloadable on platforms such as GitHub (Redmon et al. 2016).[8] Finally, although the proprietary backends of AI interface platforms are not always accessible, anyone could produce outputs with these programs. In this context, prospective controls of AI inputs would face challenges similar to those encountered in attempts to establish internet controls: they are often decentralized and fungible (Tallberg et al. 2023, 3). Arguably, AI's intangible elements could be controlled by restricting the physical basis of AI applications, as most current policies do. Hardware components establish the processing capabilities for the training and deployment of AI (Allen 2023; Scharre and Lamberth 2022). Notably, a new generation of specialized AI chips is becoming critical to training algorithms on increasingly larger data sets (Ahmed and Jenihhin 2022, 8; Scharre 2023, 42). Yet controlling the semiconductor supply chain is also not straightforward. The design and production of semiconductors is highly globalized (Thadani and Allen 2023; Mark and Roberts 2023). Simultaneously, many chips are dual-use; they are valuable in both civilian and military industries (Byrne et al. 2022). This reality renders it difficult to establish clear distinctions for technology control rules (Fist, Schneider and Heim 2023). Finally, while the training of AI models benefits significantly from edge-compute capabilities, in effect the on-system use of AI algorithms can also be facilitated by legacy processors (Shivakumar, Wessner and Howell 2023). For these reasons, some hold that the governance of military AI cannot be modelled on international regimes such as the Treaty on the Non-Proliferation of Nuclear Weapons and the Chemical Weapons Convention governing the spread of nuclear and chemical weapons, respectively (Afina and Lewis 2023).[9] Compared with these technologies, both inputs and uses of AI are harder to trace, are not strictly government run, and cannot be explicitly forbidden due to overlap between civilian and military applications (Baronchelli 2023, 2). Simultaneously,

6 For an explanation, see Ford (2018).

7 As summarized by Paul Scharre (2023) in *Four Battlegrounds*.

8 See https://github.com/ultralytics/yolov5/releases/tag/v7.0.

9 For an early view, see The Hague Centre for Strategic Studies (2018).

the innovation of AI benefits from a fragmented and open-source environment (ibid., 3). The nature of AI as a category of technologies dependent on software, hardware and people with dual-use applications requires that governance mechanisms distinguish how they apply to specific inputs.

Challenge 3: The Diversity of Actors Involved in the AI Chain

Third, a plethora of actors are involved in the production-proliferation-deployment-employment chain for AI in the military domain. These actors come from radically different backgrounds — each with their own interests and values. The design and production of AI applications is concentrated within private entities whose innovation models are motivated by favourable national (regulatory) environments. Meanwhile, AI can proliferate from both state and non-state actors (ibid.). As a result, militaries may decide to develop their own respective research and development (R&D) capabilities to maintain maximum oversight (Fischer 2022). However, production of hardware components will almost inevitably need to be outsourced. The production of semiconductors has a particularly high threshold of entry due to the technological sophistication and costs involved. This has resulted in market domination by a limited number of oligopolists remaining at the forefront of semiconductor production (J. Lewis 2022). It has also proven difficult for militaries to gain both sufficient funding and talent in-house to compete with the private sector (Krieger et al. 2021, 380). Currently, military organizations opt for software-hardware co-design in the development of AI applications (Soare, Singh and Nouwens 2023), which requires high levels of correspondence between actors involved in production (Baronchelli 2023, 3; Ekelhof 2022). In turn, this necessitates strategic choices about the development of AI applications. These choices could fuel global fragmentation and the nationalization of development processes to the detriment of transparency and openness required for building transnational frameworks for AI regulation (Tallberg et al. 2023, 3). Absent such trends, frameworks need to account for the roles and responsibilities played by different actors that face various motivations and guidelines. Effective application of norms requires clarity about where something is produced, purchased and processed, to be able to determine who is ultimately accountable when something goes wrong (von Ingersleben-Seip 2023, 800).

Challenge 4: The Difficulties of Verifying the Use of AI in the Military Domain

Fourth, it is often incredibly difficult to determine whether and how AI has been used in a military context. As such, even if rules and regulations were to be established around military AI, the process of verifying adherence to these rules would be complicated by the very nature of the technology. This is due to the way that AI is integrated into military systems as well as throughout the OODA-loop (Kwik and Van Engers 2021, 45). On the one hand, it is not always immediately clear whether a system is employing AI. Systems can be fitted with AI enhancements without necessarily appearing physically different. At the same time, these systems can switch between AI-enabled and non-AI modes or receive software updates after inspection (Scharre and Lamberth 2022). This once more raises the definitional issue of AI, specifically related to when a system can be governed. On the other hand, because AI can be applied throughout the OODA-loop, it may be difficult to tell to what extent AI influenced the outcomes of specific systems, at what points in decision making and how humans were involved (Canca 2023, 59). This challenge is already visible in modern contexts. Israel's employment of AI tools such as "The Gospel," "Lavender" and "Where's Daddy?" in Gaza has raised different questions: To what extent did AI influence or even determine target acquisition and decision making? And how were individuals prepared to operate these AI-enabled systems? (Office of the High Commissioner on Human Rights 2024; Rommen 2024). Meanwhile, on the battlefield, it has been difficult to establish the degree to which different systems employed by Azerbaijan, Russia and Ukraine are, in fact, autonomous. As such, effective governance mechanisms require greater clarity around verification issues.

Challenge 5: The Role of AI in Interstate Strategic Competition

Fifth, the intensification of interstate strategic competition in recent years has also manifested itself in the military realm. Major and non-major powers are adapting their military postures and strengthening their military capabilities, including through investments in military AI (Fischer 2022). Amid considerable uncertainty about the extent to which military AI will alter global and regional power balances, key state actors may be reluctant

to commit to governance initiatives, lest they lose perceived competitive advantages (Horowitz, Pindyck and Mahoney 2024; Horowitz 2018). This helps explain the deadlock in the deliberations of the GGE on LAWS (United Nations Office for Disarmament Affairs [UNODA] 2024; Schmitt 2022, 9). The reluctance is intensified both by the rate at which new AI applications are being developed and the existing structures, which have made investment in AI extremely lucrative (Bode et al. 2023). Moreover, military AI faces issue linkage: decisions in one area may create negative effects that transfer from one domain of interstate relations into another. States may fear that continuing to trade critical materials and components with rivals could contribute to future AI-enabled security threats. As such, globalized supply chains are more readily perceived as a risk, and states become interested in limiting each other's capacity to innovate instead of working in cooperation (Allen 2023; Palmer 2023). Overall, these considerations render it more difficult to arrive at shared positions beyond lowest-common-denominator agreements (Faesen et al. 2020, 29).

Challenge 6: The AI Power Paradox

Sixth, and finally, the current rate of AI development outpaces the rate at which policies can be formulated and adopted, creating the so-called AI power paradox (Bremmer and Suleyman 2023; Baronchelli 2023). The paradox lies in the observation that the staggering range of AI applications creates a variety of policy issues in different domains. These issues have the potential to alter the status quo in many areas of life and are therefore pressing matters to be addressed by regulations (Baronchelli 2023, 2). However, it is nearly impossible to respond to these challenges quickly precisely because of uncertainty about their impact. Governments know that policies created in the present day will contribute to path dependencies: future policies will be shaped by the decisions made today. Consequently, policies must be well thought through and evaluated. Yet, as AI applications are continually evolving, evaluations risk becoming outdated before their implementation while the need for regulation only grows. These delays may simultaneously be in the interest of certain actors, as regulations could slow down innovation. Currently, most governments worldwide lag behind private sector innovation, focusing on reactive policies.

Normative discussions about the governance of military AI need to consider both timeliness and longevity. The starting point for addressing this paradox is for states to clearly identify values that would guide the integration of AI into their militaries for responsible uses (Hashmi 2019).

International Norms on AI in the Military Domain: Taking Stock of a Crowded Landscape

International norm development for AI in the military domain is ongoing, albeit in its early stages. The blossoming of a plethora of initiatives takes place within a wider emerging regime complex of norms for AI that is horizontal and decentralized, relies primarily on soft law and involves a variety of stakeholders (Tallberg et al. 2023, 11–12). Most international initiatives targeting AI in the military domain emerged only in the last five years.[10] In part, this can be explained by the fact that AI has fast become a reality within the military domain, rather than mere futuristic speculation. These initiatives are not necessarily complementary but neither do they compete; their mandates and goals differ substantially, but there are also inevitable points where ideas overlap.

In assessing the emerging landscape, a comparative norm analysis of 13 international initiatives in the form of strategies, declarations and resolutions was conducted (see Table 2). Table 2 summarizes the reviewed initiatives from most recent to oldest, starting out with those initiatives with a military focus. Seven selected initiatives directly address AI in the military domain (see Figure 1), in addition to seven initiatives targeted at regulating AI more

10 These include international initiatives directly related to the governance of AI in the military domain, such as the REAIM Summit (2023), the United States Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy (2023), the Latin American and the Caribbean Conference on the Social and Humanitarian Impact of Autonomous Weapons (2023), the North Atlantic Treaty Organization (NATO) Artificial Intelligence Strategy (2021) and the ongoing work of the GGE on LAWS. Additionally, some states have developed their own approaches to AI in the military domain, notably Australia, France, the United Kingdom and the United States.

Table 2: Overview of 13 International Initiatives Reviewed

	Initiative	Year	Military focus?	Type of document
1	REAIM Summit Call to Action	2023	Yes	Call to action
2	Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy	2023		Declaration
3	(Draft) UNGA Resolution on Lethal Autonomous Weapons	2023		(Draft) Resolution
4	Communiqué of the Latin American and the Caribbean Conference of Social and Humanitarian Impact of Autonomous Weapons	2023		Communiqué
5	CARICOM Declaration on Autonomous Weapons Systems	2023		Declaration
6	(Summary) NATO Artificial Intelligence Strategy	2021		Strategy
7	Guiding Principles affirmed by the GGE on LAWS	2019		Report
8	(Draft) AUDA-NEPAD Artificial Intelligence Roadmap for Africa	2024	No	(Draft) Road map
9	OECD Recommendation of the Council on Artificial Intelligence	2019/ 2023		Recommendation
10	Bletchley Declaration	2023		Declaration
11	EU Artificial Intelligence Act	2023		Law
12	IEEE Position Statement on Ethical Aspects of Autonomous and Intelligent Systems	2021		Position statement
13	Charlevoix Common Vision for the Future of Artificial Intelligence	2019		Vision document

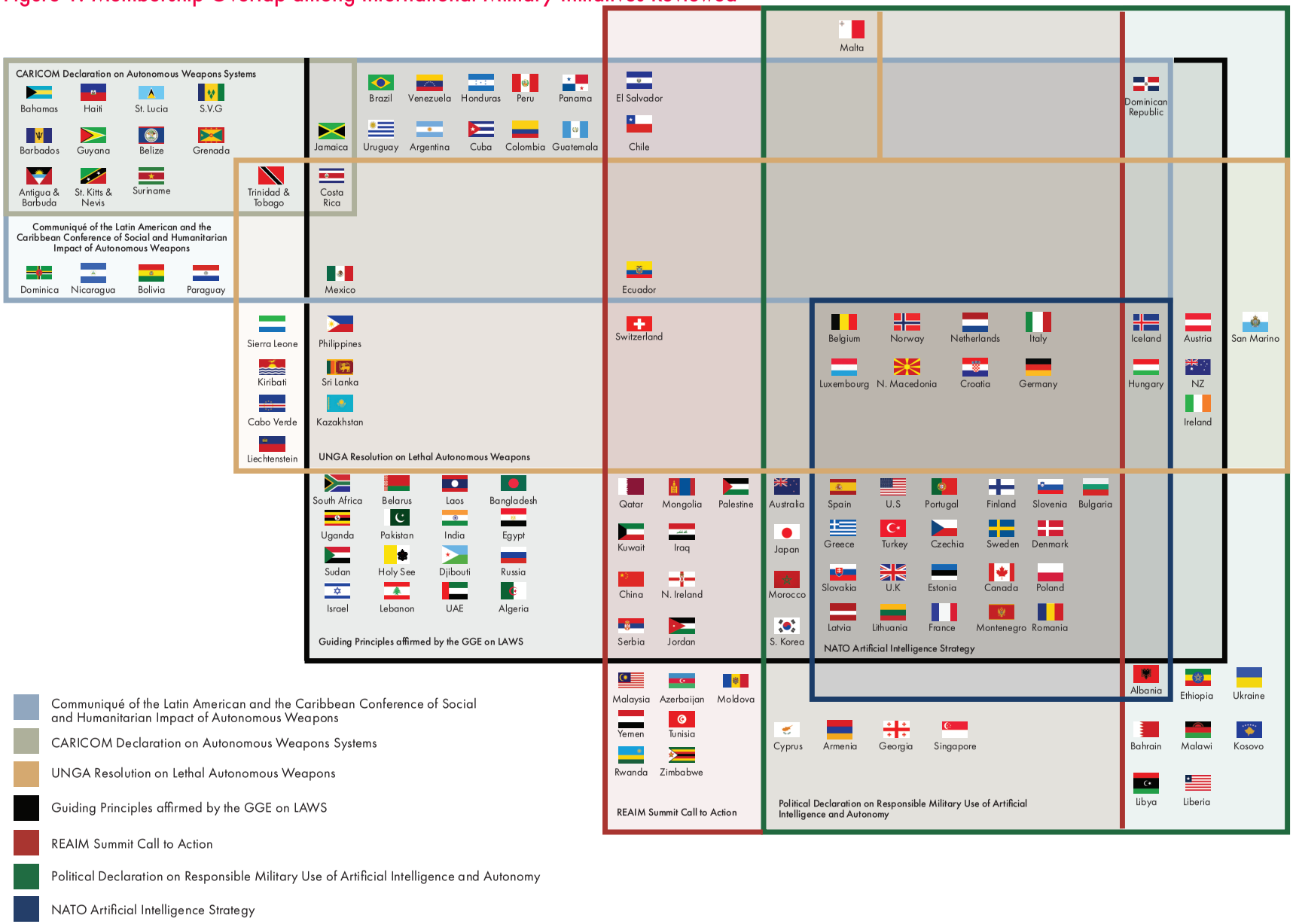
Note: UNGA = United Nations General Assembly; CARICOM = Caribbean Community; AUDA-NEPAD = African Union Development Agency-New Partnership for Africa's Development; IEEE = Institute of Electrical and Electronics Engineers.

Source: Authors.

broadly. These initiatives have significantly shaped thinking on AI norms as a whole (Schmitt 2022, 303–14; von Ingersleben-Seip 2023, 797). The dual-use nature of most AI applications makes it nearly impossible to draw a clear civil-military divide in governance. Therefore, this paper reflects on the observation that although initiatives such as the EU Artificial Intelligence Act do not apply explicitly to defence, such comprehensive regulations passed on AI, in general, inform the normative boundaries for norm development in the military domain as well.[11] Simultaneously, initiatives such as the AUDA-NEPAD Artificial Intelligence Roadmap for Africa and the OECD Recommendation of the Council on Artificial Intelligence were included as regional reflections of AI-related principles (see Figure 2).

11 Notably, the UNODA report *Towards Responsible AI in Defence: A Mapping and Comparative Analysis of AI Principles Adopted by States (2023)* takes a similar approach by reviewing all AI-related initiatives at national and international levels. The scope of this paper is more limited, with a narrower focus on primarily military-specific initiatives. See Anand and Deng (2023); von Ingersleben-Seip (2023, 802); Ekelhof (2022).

Figure 1: Membership Overlap among International Military Initiatives Reviewed



Source: Authors

Table 3: Definitions of Seven Normative Focus Areas

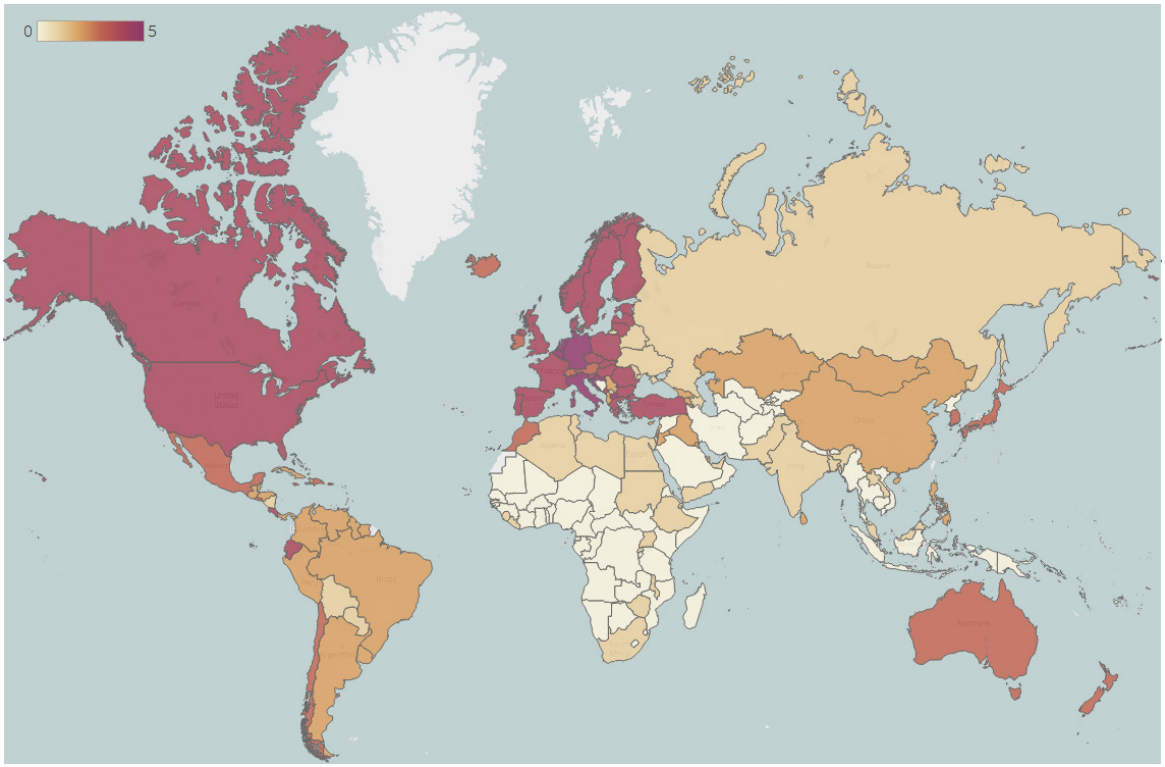
	Normative Focus Area	Definition
1	Accordance with international law	The use of AI in the military domain should be carried out in accordance with existing international humanitarian and human rights law. Any guidelines created specifically for AI in the military domain do not supersede existing international law.
2	Responsibility and accountability	Human agents maintain responsibility, and therefore accountability, for the use of AI in the military domain throughout a system's lifecycle.
3	Explainability and traceability	To maximize the benefits and minimize the risks of the use of AI in the military domain, sufficient understanding and transparency of systems, inputs and outputs is needed.
4	Bias and harm mitigation	The potential harmful consequences of biases and the general use of AI in the military domain need to be considered and addressed proactively.
5	Reliability	The use of AI in the military domain should be robust and with appropriate safeguards to ensure that systems can carry out tasks consistently and predictably.
6	Governability	Guidelines for the use of AI in the military domain should enable practitioners to detect and avoid unintended consequences as well as disengage and deactivate systems when undesired incidents occur.
7	Exchange of practices	To support the development and improvement of AI in the military domain, good and best practices should be exchanged among actors throughout system lifecycles.

Source: Authors.

A qualitative review of all the abovementioned initiatives produced seven general focus areas (see Table 3). These focus areas were deduced based on a mapping of either the explicit principles listed by the initiatives or on the values embedded in these principles. More specifically, when a numbered or lettered list was found within a document referring to principles or values, this was taken as indication of the most direct summary of the norms promoted by an initiative. If such a summary was absent, reference to norms was found in the text itself. The clustering of these categories was cross-referenced with literature analyzing the developing normative landscape around AI in the military domain. All seven normative focus areas have been independently recognized as key motifs in AI governance overall (Vestner 2022). The categories were therefore determined through iteration, taking into account existing normative categories and the saturation of specific embedded norms, in both current initiatives and wider academic and professional literature. The methodology resembles the approach adopted by Alisha Anand and Harry Deng (2023, 9), who, reviewing a broader swath of AI governance initiatives, identified 18 principles adopted by intergovernmental organizations. Some of the principles identified by Anand and Deng, but not all, overlap with the normative focus areas identified in this paper.[12]

12 Specifically, Anand and Deng identify "impartiality" and "inclusiveness" (clustered under "fairness"); "human oversight, judgement or control"; "human dignity"; "compliance with law," "data protection" and "privacy" (clustered under "lawfulness"); "proportionality"; "public engagement"; "accountability" and "responsibility" (clustered under "responsibility and accountability"); "sustainability"; "reliability," "safety" and "security" (clustered under "technical robustness"); and "explainability", "information sharing" and "traceability" (clustered under "transparency") (Anand and Deng 2023, 9). The overlap with the study is found in the "lawfulness" cluster, corresponding to what was labelled "accordance with international law" and the categories of "responsibility and accountability." Other concepts were clustered and labelled differently, but they correlate in terms of content.

Figure 2: Number of Initiatives around AI in the Military Domain Signed by States



Source: Authors (via www.mapbox.com).

The seven categories are not always strictly demarcated. Overlap exists where one category begins and another one ends, and in different disciplinary perspectives, some focus areas would not be separated.[13] Yet they are treated as discrete categories in the initiatives themselves.

For each initiative, a relative assessment was made of the extent to which attention was given to a particular normative focus area. In so doing, an admittedly rough but relatively straightforward and transparent method was used to gauge importance. Attention was assessed by looking at the number of words dedicated to a specific principle proportionate to other principles in the same document and the amount of detail used to describe it. At the same time, a semantic analysis of the content was conducted. If an initiative mentioned a point related to a normative focus area only once with little semantic elaboration, then it scored low on attention. If, proportionate to the length of the entire initiative, points related to a normative focus area were mentioned several times with some semantic elaboration, then attention was marked as medium. Finally, if most points were dedicated to one normative focus area, then attention was marked as high (see Table 4).

Based on this method, the norm analysis not only identifies key focus areas addressed by the initiatives, but also grants some insight into the extent to which thinking on focus areas has developed. If one normative focus area was consistently mentioned across initiatives but was labelled as low on attention, a gap in norm development could be inferred. Specifically, this indicates that the value of a norm was collectively recognized but that the implications and practice of the norm still lacked consensus. This serves as a first stepping stone to more in-depth analysis of the content of the norms proposed by the initiatives.

13 For example, separating "accordance with international law'' and "responsibility and accountability'' into different focus areas could obscure the fact that both responsibility and accountability are, in part, legal concepts, the tenets of which are established primarily through codified law. See Boulanin and Lewis (2023, 6).

Table 4: Assessment Framework for Normative Focus Areas

Level of Attention	Example
Low	Accordance with international law in the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy mentioned in principle B (one out of 10 points). Focus on international humanitarian law.
Medium	Responsibility and accountability in the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy attributes specific responsibilities to states, senior officials and relevant personnel, and references the need for understanding.
High	Accordance with international law in the Guiding Principles affirmed by the GGE on LAWS mentioned in principles A, C, E and H (four out of 11 points). Focus on the different ways in which lethal autonomous weapons can interact with international law.

Source: Authors.

In addition to more broadly identified advantages and disadvantages of certain norms, this paper specifically evaluates strengths and weaknesses on three dimensions: adoptability, verifiability and enforceability. This analysis is further informed by how norms are formulated and justified by initiatives themselves, what the initiatives explicitly refer to as existing avenues for further development, and assessments in the scholarly and professional literature about the efficacy of norms.

An Assessment of Strengths and Weaknesses in International Norms for AI in the Military Domain

Accordance with International Law

Accordance with international law is consistently, and often prominently, included as a key normative focus area by interstate initiatives for governing AI within and outside of the military domain. Notably, six initiatives focusing on military AI refer to international law either in the first or second point among all other listed focus areas. This is not necessarily surprising. International law is widely recognized as a necessary starting point for discussions around AI in the military domain (Vestner 2022). International humanitarian and human rights law reflects centuries of attempts to codify the ways in which polities conduct war (Sweijs 2023). As such, the existing body of international law captures areas in which the international community has reached at least some degree of agreement (Ingersleben-Seip 2023, 787). At the same time, the sometimes ambiguous formulations of international law principles provide states with leeway in the interpretation of their responsibilities.[14] Therefore, most AI governance initiatives prefer to operate within "the existing architecture," relying on proven (governance) mechanisms to address the challenges posed by AI technologies (Schmitt 2022, 305). For these reasons, it has been conceded that international law should, in one way or another, apply to AI systems and applications despite their complexity.

However, these purported strengths of the normative focus area also contribute to its weaknesses. Beyond a baseline accordance with international law, initiatives often fail to answer this question: What international law? If AI is

14 For example, the formulation of provisions of the European Court of Human Rights is vague by design to "allow for domestic contextualization" and enactment. See Ammann (2020, 179). Meanwhile, the use of undefined terms such as "responsible AI" is impacting the developing norms on AI in the military domain. See Schmitt (2022, 311).

already sufficiently regulated by international law, there needs to be clarity as to how its aspects are practically reasserted, reapplied, extended or clarified in the context of military AI (Maas and Villalobos 2023, 8; Vestner 2022). Existing initiatives do not consistently specify what kind of international law bodies, treaties or principles are relevant. Some initiatives only broadly refer to international humanitarian and human rights law, with the latter only appearing next to the former. Treaties such as the The Hague Conventions and the Geneva Convention, and potentially relevant corresponding articles such as article 36 of the Additional Protocol I of the latter,[15] are typically not mentioned. And, while presumably existing principles of international law apply, such as proportionality, precaution, distinction, military necessity and humanity (Hunter Christie et al. 2023, 4), the principles are not omnipresent.[16] Arguably, states are obligated to make assessments according to these principles as a baseline regardless of the tools used. However, there still seems to be a tension that, in many ways, reflects the challenges of international law more broadly. The fact that adherence to international law is consistently mentioned as a principle among governance initiatives indicates that states see a utility and need to reiterate that it is relevant to AI. While there is consensus that military uses of AI should not be excepted from international law as a whole, there is no agreement as to whether extant international law is sufficient (Tallberg et al. 2023, 11).

If existing international law is ultimately judged to be sufficient to govern AI in the military domain, practitioners still face the issue of interpretation. Although reliance on existing regulations alleviates some of the challenges of the AI power paradox (Bremmer and Suleyman 2023), the consistent interpretation is difficult. Even like-minded states, as among some NATO allies, "will derive their own interpretations on how principles should be best employed" (Hunter Christie et al. 2023, 13). The formulation of international laws is often, by design, a reflection of the diverging perspectives of signatories. But it also constitutes an obstacle when trying to create clarity on the acceptable issue boundaries (Goldsmith and Posner 2005; Schmitt 2022, 311). Further aggravating the problem, the lack of consensus on AI as a category of technologies overall can create complications. A key precondition of international law is that it is exercised and implemented by human agents with sufficient understanding and assessment of violations (D. A. Lewis 2023, 500). While in most areas covered by international law, these principles are practically a given, the autonomy and algorithmic nature of AI have the potential to even "defy the human-centred foundation of international law" (Garcia 2024, 27). The question of autonomy also calls for "more comprehensive research into the legal significance of mistakes of fact" in relation to fundamental principles of international humanitarian law (Pacholska 2023, 22).

In the case that states ultimately agree on the need for new legal regulations focusing specifically on AI, international deliberations would face the worst of the AI power paradox. As in the case of the EU Artificial Intelligence Act, attempts to build on existing legal frameworks with new legally binding mechanisms are complicated by rapidly developing AI capabilities. The resulting clash between legal codification and technological innovation demands a greater adaptability from regulators, which is not necessarily facilitated by existing institutions (Walters and Novak 2021). New regulation would also face the challenge of "whether to prioritize breadth of membership and inclusion or depth of mission alignment," risking either only lowest-common-denominator agreements or fractured and ineffective regimes (Maas and Villalobos 2023, 20–21).

The foundational nature of this normative focus area may entail that it is expanded upon by more detailed formulations of norms in the other areas. Therefore, this normative focus area can continue to serve as an entry point for at least one agreement: AI will not be excepted from international law. In this form, the adoptability of the principle is quite high, reflecting continued respect for international law. For norm development going forward, this lowest-common-denominator agreement may be necessary although not sufficient, especially in the fragmented landscape of AI governance. Eventually, a more comprehensive approach to clarifying relevance and applicability of international law will be needed. At the moment, norms in this focus area have low verifiability and enforceability:

15 Article 36 "provides for a specific obligation to determine, when considering the development or acquisition of a new weapons, means or method of warfare, whether its employment would, in some or all circumstances, be prohibited by any applicable rule of international law." See Boutin (2023, 145).

16 Admittedly, in part, this lack of specificity could be due to the fact that the reviewed initiative documents represent summaries of negotiations. However, this issue is recognized more broadly as well in academic publications.

there are no practical criteria for assessment nor agreement about red lines, which, if crossed, should be sanctioned. However, these two elements can be improved if this focus area is not viewed in isolation.

Responsibility and Accountability

The relationship between responsibility and accountability is both a moral and a legal one. Although "often conflated 'responsibility' and 'accountability' are distinct concepts. Accountability is scrutiny from an external point of view and is a form of 'answerability', whilst (moral) responsibility is an internal point of view, i.e. an assessment of agency" (A. Blanchard, Thomas and Taddeo 2023, 15). In the context of AI in the military domain, the ability of AI to take over otherwise human tasks raises questions about who is responsible and, consequently, accountable for actions or omissions (Boutin 2023, 141). Following Filippo Santoni de Sio and Giulio Mecacci (2021), four kinds of responsibility gaps can be distinguished in relation to military AI: the culpability gap;[17] the moral accountability gap;[18] the public accountability gap;[19] and the active responsibility gap.[20]

The importance of addressing these gaps is reflected by the fact that all but two initiatives make an explicit reference to responsibility. There is a general consensus within existing initiatives that establishing responsibility contributes to accountability mechanisms. Specifically, as Deborah G. Johnson explains, "No matter how independently, automatically, and interactively computer systems of the future behave, they will be the products (direct or indirect) of human behaviour, human social institutions, and human decision" (quoted in Pacholska 2023, 19). Ultimately, the requirements will vary depending on the goals of establishing accountability, whether it be compliance, reporting, oversight or enforcement (Novelli, Taddeo and Floridi 2023, 16). Yet this norm's practical application to military AI has several weaknesses.

In all initiatives related to military AI, there is a general reluctance to identify where the responsibility of one party begins and how accountability for undesirable outcomes would consequently be determined.[21] This pitfall is aggravated by a tendency to include buzzwords such as "meaningful human control," "human oversight" and "human responsibility," without clearly defining the requirements for *adequate* oversight and responsibility (Tigard 2021). Actors that could be held accountable are states and individuals, as well as private corporations (Pacholska 2023, 5). But the responsibility of all these actors cannot be treated in isolation. The concept of "responsible reliance" emphasizes that "natural persons involved in the development and use of an AI tool in an armed conflict need to be able to rely on what the other relevant actors did to help ensure that the tool's behaviour, performance, and effects are lawful" (Boulanin and Lewis 2023, 8). Consequently, responsibility may be better conceptualized as a process where any actor who is recognized to be involved in a chain of cumulative responsibility could be held accountable (ibid., 9).

If attribution of culpability falls only on the deployer of an AI system, "this may have a detrimental effect on the way actors involved in command and control may perceive their responsibilities" (Taddeo et al. 2021, 1716). Current formulations of responsibility and accountability can contribute to the impression that being responsible is a disadvantageous position. Not only are those "responsible" taking on risks, but they are also the ones who will bear the consequences (Santoni de Sio and Mecacci 2021, 1070). This could demotivate actors from proactively taking on responsibility to maintain perceived freedoms, creating the need for "better mechanisms to promote the moral accountability of all agents involved in the design and use of AI systems; better mechanisms of public accountability for those who design or regulate AI systems operating in the public space; and mechanisms and policies to promote a better culture of active responsibility of all the designers, managers, controllers, and users of AI systems" (ibid., 1074).

This reflects the fact that the "neat theoretical distinction between different stages of technological innovation does not always exist in practice"

17 "The risk that no human agent might be legitimately blamed or held culpable for the unwanted outcomes of actions mediated by AI systems" (Santoni de Sio and Mecacci 2021, 1059).

18 "Human agents' capacity to make sense of – and explain to each other the 'logic' of their behaviour" (ibid.).

19 Where citizens will not be "able to get an explanation for decisions taken by public agencies" (ibid.).

20 "The risk that persons designing, using, and interacting with AI may not be sufficiently aware, capable, and motivated to see and act according to their moral obligations towards the behaviour of the systems" (ibid.).

21 The reviewed initiative that most clearly stipulates limits of responsibility are the Chinese interim measures for generative AI, which place most responsibility on the supplier-side companies of generative AI services.

(La Fors, Custers and Keymolen 2019, 210). Overall, the boundaries between parties also relate to how intention and causality are assessed in the context of responsibility and accountability (Kwik and Van Engers 2021, 57). Algorithms can contribute to a so-called double fog of war, where the decisions are not only obscured by the complexity of the battlefield but also by the black-box nature of AI systems (Kwik and Van Engers 2021). At the same time, remote warfare enabled by AI physically distances actors from their actions. As such, norms should address how to identify ownership and thereby the locus of responsibility among actors.

Relatedly, existing initiatives do not explain how state and individual responsibility would be assigned. Responsibility on both levels is readily considered to be complementary and concurrent (Pacholska 2023, 7; Boutin 2023, 148). Ensuring respect for existing international law would require states to enable the individuals and organizations working with and around AI systems to consider compliance with law at all stages. Specifically, individuals who have the power to act on behalf of the state should be expected to: "(1) foresee that the effects of the system will not be unlawful (which, in turn, presupposes a sufficient understanding of the system's performance and behaviour); (2) administer the system in a manner that ensures that its operations and effects are lawful; and (3) trace the operation, performance, and effects of the system back to the relevant natural person(s) to help ensure accountability" (Boulanin and Lewis 2023, 6).

Accordingly, principles of individual criminal liability would only be applicable if explicit responsibilities, and corresponding violations, can be established. But criminal law would "be less adequate to cope with substantial shared responsibilities derived from manifold individual small faults" (Santoni de Sio and Mecacci 2021, 1074). As such, "upholding the responsibility of collective actors such as states acknowledges the structural forces that drive the development and use of AI" (Boutin 2023, 134). State responsibility entails the formulation of protocols and procedures to prevent any misuse or abuse, to ensure the cessation of wrongful activity and the provision of reparations, and the assurance that measures will be taken to prevent future incidents.

For norms around responsibility and accountability to be effective, more careful consideration is required for the different types of responsibility attributable to all relevant actors within an AI life cycle, the different expectations at the state and individual level, and the designation of ownership. As responsibility and accountability touch at the very core of what AI systems offer for the military domain, the adoptability of this normative focus area is relatively high. However, until stricter actor-specific criteria are determined, verifiability and enforceability of this norm remain difficult. One potential avenue would be to look at responsibility within private organizations to see how these principles are integrated into design and technical standards. The international community can step in when red lines of responsibility are crossed, in line with international law.

Explainability and Traceability

Existing initiatives emphasize that there should be sufficient understanding of the workings of an AI system by relevant actors. As such, this norm category is closely related to both the focus areas of responsibility and accountability as well as reliability. Explainability in AI "refers to the ability to provide a semantic expression (as opposed to merely quantitative and operational) to why decision processes developed in a certain way," while traceability can be understood as instances in which "certain outputs from an AI algorithm can be traced to certain inputs, as if going back in the decision chain" (Hunter Christie et al. 2023, 7–9). Both concepts are related to transparency, which is the extent to which information about a system or its development is accessible by stakeholders (A. Blanchard, Thomas and Taddeo 2023, 19).[22] Existing initiatives around AI in the military domain recognize that explainability and traceability are related, and agree that these two elements are pivotal to ensuring the responsible use of AI.

Still, principles within this normative focus area could benefit from further explicitation. Specifications of how much explainability and/or traceability is required should be established for different types of applications. Insufficient visibility of the relationship between inputs and outputs in certain AI applications could feed into the risk of — another — double black box, where technical ambiguity enhances military secrecy. In essence, lack of understanding of the workings of an AI system could be motivated by the desire to secure

22 The relationship between the three concepts can be summarized as follows: "Traceability is necessary, but not sufficient for explainability. Explainability is necessary, but not sufficient for full transparency." See Hunter Christie et al. (2023, 9).

sensitive information (D. A. Lewis 2023). However, in some situations, a lack of explainability and/or traceability could enhance AI performance. Black boxes can yield higher performance as, at least for now, "there is [a] clear trade-off between the performance of a machine learning model and its ability to produce explainable and interpretable predictions" (Linardatos, Papastefanopoulos and Kotsiantis 2020, 18). An acceptable balance needs to be struck between the dual imperatives of international law and military competitiveness, as articulated by Jonathan Kwik and Tom Van Engers: "A shift to less transparent AI is a corollary of the military needs and circumstances to which such models can provide a solution....To benefit optimally from AI technology, performance must be maximised, but only to the extent that the product remains within the constraints of the law" (Kwik and Van Engers 2021, 44).

Without elaboration, it remains unclear to what extent explainability and traceability are desirable for different AI systems. The utility of explainability and traceability measures in systems would also be dependent on who would have access to the resulting information and how they could benefit from it. Without some transparency, where relevant decision makers and individuals in positions of power would have access to information, traceability itself would be of little use (Taddeo et al. 2021, 1718). Even if all relevant practitioners had access to fully explainable and/or traceable AI, only a small number of them would be able to interpret it without comprehensive reskilling. A study by Michael Horowitz and Lauren Kahn found that "those with the lowest level of experience with AI are slightly more likely to be algorithm-averse, [therefore] automation bias occurs at lower levels of knowledge before levelling off as a respondent's AI background reaches the highest levels" (Horowitz and Kahn 2024), supporting the argument for additional education to prevent bias-based issues. As AI has the potential to influence the entirety of the OODA-loop, this could necessitate additional education for every aspect of military AI. As such, norms around explainability and traceability may have greater institutional implications. Albeit analyzing a national initiative (the US Department of Defense AI principles), Alexander Blanchard, Chris Thomas and Mariarosaria Taddeo find that the existing guidelines "delineate the institutional attitude towards the adoption of AI, but they do not offer specific guidance to address the problems that may emerge in applying the principles to specific cases....This means that responsibility for making complex ethical assessments is devolved onto practitioners who may lack the necessary expertise" (A. Blanchard, Thomas and Taddeo 2023, 8).

Finally, while the discourse about norm development generally eschews technical details, engagement with the technical feasibility of the implementation of norms is crucial. Responsibility is not just about who pays for mistakes, but also about being involved and transparent from the beginning with all actors. When it comes to the verifiability and enforceability of explainability and traceability, this focus area has perhaps some of the most potential as "models can be audited in multiple ways, ranging from internal code and training process reviews to fuzzing and deterministic testing, and different applications will require different degrees of capability auditing" (Dunnmon et al. 2021, 29). Even with intrusive inspections, concerns over potentially exposed vulnerabilities could be assuaged by installing privacy-preserving software verification and minimal external monitoring functions (Scharre and Lamberth 2022). At the same time, leaving the specifications of thresholds to technical experts may also worsen the culpability gap where "technical experts may (honestly) believe that nobody is to blame for an accident because they have done what could reasonably be expected from them" (Santoni de Sio and Mecacci 2021, 1071).

The normative focus area of explainability and traceability is one that has the potential to perform relatively well across all three criteria. Adoptability of related norms is medium because it is generally deemed desirable to understand the technology, especially because of how understanding then relates to accountability. Verifiability is also feasible as a system can be checked and measured, for example, for whether self-explaining mechanisms have been built in. Still, measuring explainability itself remains difficult as it is context- and effect-based. Enforceability is complicated, although progress could be made by focusing standards and protocols in the production of systems. The main challenges then are how to enhance transparency within a secretive environment that would not create strategic disadvantages for actors who adhere to norms more closely, and how to create sufficient distinction in the requirements for different AI applications regarding measures for explainability, traceability and/or transparency.

Bias and Harm Mitigation

Bias and harm mitigation appears in all initiatives, except for the draft UNGA resolution on LAWS, and receives ample attention. Normative principles around bias and harm mitigation rely on matured conceptualizations of the potential risks associated with bias. Notably, bias itself is seen as a type of harm as it can potentially lead to issues ranging from discrimination and inequity to unexpected emergent behaviours in systems (Gray et al. 2024, 688). Five main sources of bias can be distinguished following Magnus Gray et al. (2024), which derive from uses in the civil domain but equally apply to the military domain:

- research designs that reflect biases of developers;
- training data that replicates incomplete or non-representative samples;
- input representations that capture "societal attitudes and display semantic biases";
- model architectures that may "[compound or amplify] existing inequities"; and
- real-world uses that follow biased applications. (ibid., 688)

Reflecting this understanding, the EU Artificial Intelligence Act devotes particular attention to the importance of diverse and comprehensive data sets in lowering the likelihood of bias in outputs, which is equally relevant to social as it is to physical input data.

Still, this norm category features a number of weaknesses. First, this focus area faces "normative creep" where a variety of different concepts, such as human rights, diversity, equity and harm, have been intertwined, without clear prioritization. This has resulted in the convolution of different dimensions that will need to be pulled apart to be addressed in practice. Second, the suggested solutions fall short of dealing with the identified problems associated with bias and harm. Although awareness of the roots of bias formation is developing, current initiatives fail to go one level deeper and offer guidelines for acquiring diverse and trustworthy data and other inputs. Third, initiatives typically fail to define what constitutes actual harm, requiring reference to concrete principles of human rights frameworks (Ams 2023). Fourth, the formulation of principles around bias and harm, which, despite stressing mitigation, in fact focus more on responses to harmful consequences instead of prevention. In part, this may be because some harmful effects cannot be predicted. However, often especially in a military context, responses to harm will come too late (Maas and Villalobos 2023, 53). This is a broader critique of "after-the-fact legal accountability… — even in the most robust and efficient legal regimes, anyone would far prefer to have prevented a harm in the first place than to be eventually compensated for it" (Crootof 2024).

Simultaneously, some types of bias and harm, which academics identify as potentially stemming from military AI, are not mentioned by governance initiatives (for example, imbuing machines with ethical faculties). In this context, "terms like 'ethical,' 'intelligent,' and 'responsible' in the context of machines can lead to false attributions and mythical tropes implying that inanimate AI agents are capable of moral reasoning, compassion, empathy, mercy, etc. and thus might perform more ethically and humanely than humans in warfare" (J. Johnson 2024, 74).

The anthropomorphizing of AI is likely to generate false expectations and obfuscates how AI impacts human decision-making processes, intensifying automation bias. Consequently, "the design of AI agents for hybrid teaming must embody both the positive and potentially negative psychological implications of anthropomorphism" (ibid., 73), including dehumanization, groupthink and diffused moral culpability (Santoni de Sio and Mecacci 2021).

Overall, the adoptability of bias and harm mitigation is comparatively high with many states signalling openness. Further norm elaboration and refinement will also help address many of the weaknesses discussed above. Once addressed, this normative focus area may also be highly amenable to verifiability and, albeit to a lesser extent, enforceability. Cases of failure with known AI systems can be observed; however, the issue of establishing AI-integration remains an obstacle. Finally, it is possible to formulate protocols and standards, both at the technical level and at the operational level, to address risks associated with bias and harm, by defining a set of mission-specific properties, standards and requirements for systems (Hoffman and Kim 2023, 22–24).

Reliability

Reliability can be defined as "the probability that a system or product will perform in a satisfactory manner for a given period of time when used under specified operating conditions" (B. S. Blanchard and Blyler 2016, 144). In a military context, reliability hinges on predictability and controllability for end users: "Operators must be able to predict with a high degree of accuracy how the weapon will behave after being deployed. A weapon would not be adequately controllable, and therefore unlawful, if there is more than a remote possibility that it could perform in an unforeseeable way" (Kwik and Van Engers 2021, 53).

An additional challenge that comes with AI applications in the military domain is that battlefields are characterized by friction and fog. Because military AI applications interact with a complex and unpredictable environment, "all autonomous systems exhibit a degree of inherent operational unpredictability, even if they do not fail or the outcomes of their individual action can be reasonably anticipated" (Holland Michel 2020, 5). Reliability can thus only be meaningfully gauged when a system's performance can be evaluated between the past and present. Specifically, a system can be checked for consistency of "how often and for how long the outputs of a system are correct; and whether the system can scale up to elaborate data that diverge from training and test data" (Taddeo et al. 2022, 12).

One weakness in existing initiatives is the lack of specification of the (in)appropriate consequences of AI use with and how they relate to specific parameters and contexts. Notably, greater autonomy of systems and more dynamic environments breed a higher chance of unpredictable, emergent behaviour (Trusilo 2023, 5). This does not necessarily mean that unpredictability and reliability are on opposite sides of the spectrum. In fact, practitioners should "confront the possibility that behaviour that is innovative but less predictable can lead to increasing reliability" (ibid., 4). This can, in turn, generate new ethical problems, for example: "Opponents of robotic swarm technology may argue that unpredictability at the micro level means there is no longer the required level of explainability or transparency.... In contrast, proponents of such a system may argue that increased reliability and robustness at the macro level make a swarm system the logical choice for real-world conflicts" (Trusilo 2022).

The next step for norm development in this focus area would therefore be to draw up specifications for different levels of assessment for reliability informed by the unique requirements of systems with distinct tasks. It is, for example, reasonable to assume that systems more embedded in decision-making processes would have more stringent reliability requirements (Taddeo et al. 2022, 18; Boulanin and Lewis 2023, 8). In addition, it would require the formulation of a layered ethical framework that stipulates conditions of reliability at these different levels of operations.

Still, the adoptability of this normative focus area remains comparatively high, with only two of the reviewed initiatives not making any reference to it. Reliability itself is a basic requirement for all weapons systems. Yet unpredictability does not always lead to increased risk. Reliability can therefore become a more robust norm if different types of risk are categorized as more or less acceptable and more or less predictable, "leading to a 'meta-level' of overall risk" (Taddeo et al. 2022, 35). In light of these considerations, the way that AI models and systems are tested can be adjusted to be dynamic and iterative, ultimately being able to account for uncertainty (Trusilo 2023, 11). Embedding technically enabled measures of reliability could, in turn, "reduce the need for costly physical enforcement (or threats thereof in order to deter certain actions)" (Sastry et al. 2024, 55).

Governability

In existing initiatives, governability establishes the need to ensure AI-enabled systems are configured in such a way that humans can take control whenever deemed necessary (Oniani et al. 2023, 225). It includes immediate response mechanisms for disengagement and deactivation in the case of unintended or unaccounted-for consequences. The "ability of human agents to contest and override AI decisions, when these should be considered mistaken or inappropriate" (A. Blanchard, Thomas and Taddeo 2023, 1719), is especially relevant for unknown unknowns, or situations that people are neither aware of nor understand (Baronchelli 2023). Yet the added utility of governability as a normative focus area is not readily apparent because issues of control and understanding are reflected in other areas as well. Still, it is prominently referenced in the US Political Declaration and the NATO principles.

There are a number of weaknesses that need to be addressed for governability to evolve into a

mature norm. First, challenges associated with responsibility and accountability would need to be resolved. This requires dealing with the plethora of actors involved in the production of AI systems and identifying the locus of responsibility to intervene to disengage and deactivate a system.

In addition, sharing governability mechanisms for specific systems between states also runs into challenges: "When one nation lends an AI capability to another nation, assurances are likely to be required as to the ethicality (and legality) of that capability given the varying ethical cultures and legal frameworks under which it was designed and developed" (A. Blanchard, Thomas and Taddeo 2023, 11).

The question presents itself as to which governance guidelines to adhere to. These tensions reveal themselves in ongoing debates where China's tighter concept of LAWS "contrasts with the approaches of other parties within the 'ban group' that prefer to identify LAWS based on characterising aspects such as autonomy and human control" (Bode et al. 2023, 5). The adoptability of this norm is therefore relatively low, its verifiability comparatively higher, but its enforceability again low.

This problem of transference also applies to non-state actors, even if they are not signatories to international norms (Tallberg et al. 2023). The decentralized and open-source nature of some AI algorithms means that a ream of applications is available to and can be proliferated by non-state actors. Public-private collaboration is therefore "key in incorporating software restrictions on commercial robotics, for example, which would address the potential consequences" of such access. This consideration is, in part, addressed in the normative focus area on reliability, which emphasizes precautions against hijacking and precaution, but considerations of harm should also go beyond the individual. A systems approach not only benefits bias and harm mitigation, but also addresses issues related to responsibility and accountability, because of "second- and third-order effects of the use of AI in various phases of operations" (Azafrani and Gupta 2023, 27).

Exchange of Practices

The final normative focus area identified is the exchange of practices, most prominently stressed by the Charlevoix Common Vision for the Future of Artificial Intelligence, but also the REAIM Summit Call to Action. Within this category, there is an emphasis on the notion that there are various actors involved with AI systems throughout the life cycle that could benefit from being mutually informed about R&D (van Hooft, Boswinkel and Sweijs 2022; von Ingersleben-Seip 2023, 789). This reflects an awareness that information exchange through multi-stakeholder engagement is necessary in a field where a large share of technological advancements take place outside of conventional, state-linked institutions. In turn, exchanges of accurate information and confidence-building measures about the state of the art of technologies could manage expectations between rivals, prevent arms races due to perceived threats and improve the safety practices for systems globally (Horowitz, Kahn and Mahoney 2020).

Despite this, elaboration of which actors should be involved in the formation of guidelines for AI systems is generally absent (Tallberg et al. 2023, 20; A. Blanchard, Thomas and Taddeo 2023, 13). It is rarely mentioned that principles designed to limit harm should not be an impediment to peaceful use or innovation of AI technologies (von Ingersleben-Seip 2023, 289). Additionally, the overall idealistic tone within this norm category fails to address the inherent risk of conflicting interests, such as motives within the private sector and threats of rivals both within and between states. In the current climate of great-power competition, it is unlikely that the exchange of practices will be truly global or sufficiently deep to achieve the desired effects. In discussing different governance regimes, Matthijs Maas and José Jaime Villalobos (2023, 24) observe that rival governments are less willing to work together due to concerns over security and proliferation, impeding the willingness of technical experts to collaborate. While international standard-setting organizations have an important role in shaping the overall discussion and development of AI governance, many are not concerned with military-specific initiatives (Schmitt 2022, 311; von Ingersleben-Seip 2023). Compared to civilian data-sharing accords, the exchange of information related to AI systems in the military domain would involve data that is confidential, with direct bearing on security (Trabucco and Maas 2023, 10). The additional security around AI systems is motivated, at least in part, by risks of poisoning or spoofing (ibid., 11). As such, it is more likely that exchanges of practices will splinter across blocs. In a way, this is already occurring: the establishment of the AI Partnership

for Defense in 2020; the AUKUS (Australia, United Kingdom and United States) security pact in 2022 wherein Pillar II focuses on emerging technologies including AI; and the collaboration between China and Russia on AI reflect adherence to the norm, but only in limited groups (ibid., 3). Meanwhile, in international fora such as the GGE on LAWS, Russia and the United States continue to oppose any measures aimed at controlling development (Bode et al., 2023, 6). Such smaller gatherings aimed at exchange of best practices are not sufficiently inclusive, with key actors being underrepresented (Stanley-Lockman 2021, 2).

Yet even within like-minded blocs, there may be issues related to exchanges of practices. Within NATO, sufficient steps need to be taken to ensure the interoperability of systems and infrastructures, which would "[require] adequate and potentially continuous data sharing" (Trabucco and Maas 2023, 10). For sufficient assurances of safe and verifiable information sharing, "greater research and investment could help increase visibility into AI capabilities, development, and deployment, and thus make strong international agreements on AI viable" (Sastry et al. 2024, 42). Even within tighter-knit groupings such as the European Union, there has been a baseline "inconsistency between the European Commission's position on excluding military AI from its emerging AI policy, and at the same time EU policy initiatives targeted at supporting military and defence elements of AI on the EU level," raising questions as to the consistency of practices within the European Union itself (Lingevicius 2023, 18).

Overall, this normative focus area is underdeveloped, which is reflected in its minimal inclusion across the initiatives. Additional clarifications for the types of practices that should or could be prioritized for exchanges include: "joint tests, trials, experimentation, training, exercises, and modelling and simulation [, and] using defence [science and technology] agreements to cooperate on shared [R&D] priorities [to] build good will for other forms of AI cooperation, including alignment with democratic values [and] technical, human, and procedural measures that foster policy and personnel…to [advance] interoperable AI adoption" (Stanley-Lockman 2021, 2).

Conclusions and Recommendations

Many military applications of AI that were considered science fiction only a few years ago have started to materialize in today's world. More are yet to come. Following this trend, governments have started to formulate norms to regulate the use of AI in the military domain. Thus far, they have been trailing AI developments rather than shaping them. A comprehensive agenda for the development of norms in this wider sphere lies beyond the scope of this paper. The following conclusions and recommendations are in place.

Norm development efforts face considerable challenges related to the breadth of AI as an all-purpose technology, the diversity of actors involved in the AI life cycle and the variety of inputs involved in the creation of AI applications, the difficulty of ascertaining military use of AI, competitive dynamics associated with AI given its crucial role in perceived advantages in interstate competition, and the so-called AI power paradox, where the pace of AI development exceeds the rate of policy formulation and adoption. These challenges make it necessary to consider adoptability — based on the interests and values of key actors — verifiability and enforceability, in alignment with core principles as enshrined in international law, when discussing norms for AI in the military domain.

Because of path dependency, future policies will be shaped by the decisions made in the present. The starting point for addressing the AI power paradox is to first clarify the overall *values*, and respective red lines, to maintain in general AI-related policies and then to proceed iteratively applying it to more specific use cases. It is crucial to start doing so now. Strategic competition in combination with epistemological uncertainty will inevitably put a strain on international efforts to find consensus. It also provides a push toward lowest-common-denominator agreements. However, this should not discourage normative efforts. Historically, laying the foundations first, for subsequent normative efforts to build on, has proven to be conducive to finding agreement later.

International norm discussions in this sphere are already taking place in a relatively crowded space. A variety of international initiatives have been

launched in recent years that inevitably overlap in themes and substance. Most of these initiatives still find themselves concerned with rather high-level conceptual issues, eschewing technical detail and lacking concrete operationalization. It is therefore important to leave the conceptual plane and delve into technical and operational specifics. As the analysis in this paper shows, the good news is that there are plenty of opportunities for the further development of norms, taking into account adoptability, verifiability and enforceability. The detailed examination of normative focus areas presented in this paper offers specific levers for further normative development along these three dimensions, which include the following recommendations:

- **Accordance with international law:** focus on articulating and specifying which core international law principles and bodies are relevant to which uses of AI in the military domain.
- **Responsibility and accountability:** consider the entire AI life cycle and production-proliferation-deployment-employment chain; clarify the onus of responsibility, including by clearly demarcating areas of responsibility and distinguishing between individual and state responsibility.
- **Explainability and traceability:** make sure technical details are part of political deliberations; address the "black box" by taking into account both hard- and software specifications as well as rules and procedures.
- **Bias and harm mitigation:** create standards and protocols defining the nature of responsible practices; target prevention efforts *ex ante* rather than only *ex post*.
- **Reliability:** create standards and protocols to assess how a system's reliability can be evaluated; engage governments and industries to use these.
- **Governability:** identify where the locus of responsibility lies to intervene in what part of the chain and clarify how national and international governance initiatives relate to each other.
- **Exchange of practices:** promote information exchange through multi-stakeholder engagement; develop confidence-building measures; address conflicting incentives of private and public actors.

Because norm development in this sphere is at its early stages, the sheer amount of attention and energy as well as the diversity of initiatives can be seen as an opportunity: it can help boost momentum for further norm development. Complementary efforts can amplify and inform each other. The assortment of initiatives can help spur development along before more detailed specifications of higher-level norms will land in specialized agreements and treaties. The role here is to complement and amplify. Together, initiatives can start to form an institutionalized regime of norms, rules and regulations guiding state behaviour.

In the end, the multifaceted nature of AI requires a multipronged approach. As such, the priorities for an agenda for norms development in this sphere should focus on formulating norms that steer the development and use of AI in the military domain toward an optimal trade-off between maximizing benefits and minimizing risks while adhering to fundamental ethical principles.

Acknowledgements

The authors would like to express their gratitude to Ayla Elzinga for her considerable research assistance and Julia Döll for creating the visuals. The authors also would like to thank two anonymous reviewers and Branka Marijan from CIGI for their valuable feedback, which has strengthened the paper's quality. The responsibility for the final result lies with the authors.

Works Cited

Afina, Yasmin and Patricia Lewis. 2023. "The nuclear governance model won't work for AI." Chatham House, June 28. www.chathamhouse.org/2023/06/nuclear-governance-model-wont-work-ai.

Ahmed, Foisal and Maksim Jenihhin. 2022. "A Survey on UAV Computing Platforms: A Hardware Reliability Perspective." *Sensors* 22 (16): 1–25. https://doi.org/10.3390/s22166286.

Allen, Gregory C. 2023. "Blocking China's Access to AI Chips Matters to U.S. National Security." Center for Strategic and International Studies, July 31. www.csis.org/analysis/blocking-chinas-access-ai-chips-matters-us-national-security.

Ammann, Odile. 2020. *Domestic Courts and the Interpretation of International Law*. Leiden, The Netherlands: Brill Nijhoff.

Ams, Shama. 2023. "Blurred lines: the convergence of military and civilian uses of AI & data use and its impact on liberal democracy." *International Politics* 60 (4): 879–96. https://doi.org/10.1057/s41311-021-00351-y.

Anand, Alisha and Harry Deng. 2023. "Towards Responsible AI in Defence: A Mapping and Comparative Analysis of AI Principles Adopted by States." Research Brief. Geneva, Switzerland: UNIDIR. https://unidir.org/publication/towards-responsible-ai-in-defence-a-mapping-and-comparative-analysis-of-ai-principles-adopted-by-states/.

Argomaniz, Javier. 2010. "When the EU is the 'Norm-taker': The Passenger Name Records Agreement and the EU's Internalization of US Border Security Norms." In *The External Dimension of Justice and Home Affairs: A Different Security Agenda for the European Union?*, edited by Sarah Wolff, Nicole Wichmann and Gregory Mounier, 117–34. Abingdon, UK: Routledge.

Azafrani, Rachel and Abhishek Gupta. 2023. "Bridging the civilian-military divide in responsible AI principles and practices." *Ethics and Information Technology* 25 (2). https://doi.org/10.1007/s10676-023-09693-y.

Baronchelli, Andrea. 2023. "Shaping New Norms for AI." *arXiv*, July 17. https://doi.org/10.48550/arXiv.2307.08564.

Blanchard, Alexander, Chris Thomas and Mariarosaria Taddeo. 2023. "Ethical Governance of Artificial Intelligence for Defence: Normative Tradeoffs for Principle to Practice Guidance." SSRN. https://doi.org/10.2139/ssrn.4517701.

Blanchard, Benjamin S. and John E. Blyler. 2016. *System Engineering Management.* 5th ed. Hoboken, NJ: Wiley. http://archive.org/details/systemengineerin0000blan_j3e2.

Bode, Ingvild. 2023. "Contesting Use of Force Norms Through Technological Practices." *Heidelberg Journal of International Law* 83 (1): 39–64. https://doi.org/10.17104/0044-2348-2023-1-39.

Bode, Ingvild, Hendrik Huelss, Anna Nadibaidze, Guangyu Qiao-Franco and Tom F. A. Watts. 2023. "Prospects for the global governance of autonomous weapons: comparing Chinese, Russian, and US Practices." *Ethics and Information Technology* 25 (1): 1–15. https://doi.org/10.1007/s10676-023-09678-x.

Boulanin, Vincent and Dustin A. Lewis. 2023. "Responsible Reliance Concerning Development and Use of AI in the Military Domain." *Ethics and Information Technology* 25 (1). https://doi.org/10.1007/s10676-023-09691-0.

Boutin, Bérénice. 2023. "State responsibility in relation to military applications of artificial intelligence." *Leiden Journal of International Law* 36 (1): 133–50. https://doi.org/10.1017/S0922156522000607.

Bremmer, Ian and Mustafa Suleyman. 2023. "The AI Power Paradox." *Foreign Affairs*, August 16. www.foreignaffairs.com/world/artificial-intelligence-power-paradox.

Byrne, James, Gary Somerville, Joe Byrne, Jack Watling, Nick Reynolds and Jane Baker. 2022. *Silicon Lifeline: Western Electronics at the Heart of Russia's War Machine*. London, UK: Royal United Services Institute for Defence and Security Studies. August. https://rusi.org/explore-our-research/publications/special-resources/silicon-lifeline-western-electronics-heart-russias-war-machine.

Canca, Cansu. 2023. "AI and Governance in Defence Innovation: Implementing an AI Ethics Framework." In *The AI Wave in Defence Innovation: Assessing Military Artificial Intelligence Strategies, Capabilities, and Trajectories*, edited by Michael Raska and Richard A. Bitzinger. New York, NY: Routledge.

Cervasio, Chiara, Nicholas J. Wheeler and Mhairi McClafferty. 2024. *Report: Crisis Prevention and Management in South Asia: Mutual Confidence, Risk, and Responsibility*. London, UK: British American Security Information Council. https://basicint.org/report-crisis-prevention-and-management-in-south-asia/.

Cohen, Stanley. 2001. *States of Denial: Knowing about Atrocities and Suffering*. Cambridge, UK: Polity.

Cohen, Raphael S., Elina Treyger, Nathan Beauchamp-Mustafaga, Asha Clark, Kit Conn, Scott W. Harold, Michelle Grisé et al. 2023. "Little in Common: Prospects for U.S.-China and U.S.-Russia Security Cooperation." Research Brief. RAND Corporation, February 20. www.rand.org/pubs/research_briefs/RBA597-1.html.

Crootof, Rebecca. 2024. "Symposium on Military AI and the Law of Armed Conflict: Front- and Back-End Accountability for Military AI." *Opinio Juris* (blog), April 2. https://opiniojuris.org/2024/04/02/symposium-on-military-ai-and-the-law-of-armed-conflict-front-and-back-end-accountability-for-military-ai/.

Dastani, Mehdi, Paolo Torroni and Neil Yorke-Smith. 2018. "Monitoring norms: a multi-disciplinary perspective." *Knowledge Engineering Review* 33 (e25): 1–22. https://doi.org/10.1017/S0269888918000267.

Dixon, Jennifer M. 2017. "Rhetorical Adaptation and Resistance to International Norms." *Perspectives on Politics* 15 (1): 83–99. https://doi.org/10.1017/S153759271600414X.

Drexel, Bill and Michael Depp. 2023. "Every Country Is on Its Own on AI." *Foreign Policy*, June 13. https://foreignpolicy.com/2023/06/13/ai-regulation-international-nuclear/.

Dunnmon, Jared, Bryce Goodman, Peter Kirechu, Carol Smith and Alexandrea Van Deusen. 2021. "Responsible AI Guidelines in Practice." Defense Innovation Unit, November 10.

Ekelhof, Merel. 2022. "Responsible AI Symposium – Translating AI Ethical Principles into Practice: The U.S. DoD Approach to Responsible AI." Lieber Institute West Point, November 23. https://lieber.westpoint.edu/translating-ai-ethical-principles-into-practice-us-dod-approach/.

Faesen, Louk, Alexander Klimburg, Simon van Hoeve and Tim Sweijs. 2021. *Red Lines & Baselines: Towards a European Multistakeholder Approach to Counter Disinformation*. The Hague, The Netherlands: The Hague Centre for Strategic Studies. October. https://hcss.nl/report/red-lines-baselines/.

Faesen, Louk, Tim Sweijs, Alexander Klimburg, Conor MacNamara and Michael Mazarr. 2020. *From Blurred Lines to Red Lines: How Countermeasures and Norms Shape Hybrid Conflict*. The Hague Centre for Strategic Studies Paper Series. The Hague, The Netherlands: The Hague Centre for Strategic Studies. https://hcss.nl/news/new-report-from-blurred-lines-to-red-lines-countermeasures-and-norms-in-hybrid-conflict/.

Farrell, Theo. 2005. "World Culture and Military Power." *Security Studies* 14 (3): 448–88. https://doi.org/10.1080/09636410500323187.

Finnemore, Martha and Kathryn Sikkink. 1998. "International Norm Dynamics and Political Change." *International Organization* 52 (4): 887–917. https://doi.org/10.1162/002081898550789.

Fischer, Sophie-Charlotte. 2022. "Military AI Applications: A Cross-Country Comparison of Emerging Capabilities." In *Armament, Arms Control and Artificial Intelligence: The Janus-faced Nature of Machine Learning in the Military Realm*, edited by Thomas Reinhold and Niklas Schörnig, 39–55. Cham, Switzerland: Springer International. https://doi.org/10.1007/978-3-031-11043-6_4.

Fist, Tim, Jordan Schneider and Lennart Heim. 2023. "Chinese Firms Are Evading Chip Controls." Center for a New American Security, June 21. www.cnas.org/publications/commentary/chinese-firms-are-evading-chip-controls.

Florini, Ann. 1996. "The Evolution of International Norms." *International Studies Quarterly* 40 (3): 363–89. https://doi.org/10.2307/2600716.

Ford, Martin. 2018. *Architects of Intelligence: The Truth about AI from the People Building It*. Birmingham, UK: Packt.

Garcia, Denise. 2023. *The AI Military Race: Common Good Governance in the Age of Artificial Intelligence*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780192864604.001.0001.

———. 2024. "Algorithms and Decision-Making in Military Artificial Intelligence." *Global Society* 38 (1): 24–33. https://doi.org/10.1080/13600826.2023.2273484.

Global Commission on the Stability of Cyberspace. 2019. *Advancing Cyberstability*. Final Report. The Hague, The Netherlands: The Hague Centre for Strategic Studies. November. https://hcss.nl/report/advancing-cyberstability-final-report/.

Goldsmith, Jack L. and Eric A. Posner. 2005. *The Limits of International Law*. Oxford, UK: Oxford University Press.

Gould, Lauren, Linde Arentze and Marijn Hoijtink. 2024. "Assembling the Future of Warfare: Innovating Swarm Technology within the Dutch Military-Industrial-Commercial Complex." In *Beyond Ukraine: Debating the Future of War*, edited by Tim Sweijs and Jeffrey H. Michaels. New York, NY: Oxford University Press.

Goussac, Netta, Natalia Jevglevskaja, Rain Liivoja and Lauren Sanders. 2023. *Enhancing the Legal Review of Autonomous Weapon Systems: Report of an Expert Meeting*. Brisbane, Australia: Law and the Future of War Research Group, TC Beirne School of Law, University of Queensland. https://doi.org/10.14264/2bbfd31.

Gray, Magnus, Ravi Samala, Qi Liu, Denny Skiles, Joshua Xu, Weida Tong and Leihong Wu. 2024. "Measurement and Mitigation of Bias in Artificial Intelligence: A Narrative Literature Review for Regulatory Science." Clinical Pharmacology & Therapeutics 115 (4): 687–97. https://doi.org/10.1002/cpt.3117.

Gurowitz, Amy. 2006. "The Diffusion of International Norms: Why Identity Matters." *International Politics* 43 (3): 305–41. https://doi.org/10.1057/palgrave.ip.8800145.

Hashmi, Ali. 2019. "AI Ethics: The Next Big Thing In Government: Anticipating the impact of AI Ethics within the Public Sector." World Government Summit 2019. February. www.worldgovernmentsummit.org/observer/reports/2019/detail/ai-ethics-the-next-big-thing-in-government.

Hoffman, Wyatt and Heeu Millie Kim. 2023. "Reducing the Risks of Artificial Intelligence for Military Decision Advantage." Policy Brief. Center for Security and Emerging Technology. March. https://cset.georgetown.edu/publication/reducing-the-risks-of-artificial-intelligence-for-military-decision-advantage/.

Holland Michel, Arthur. 2020. *The Black Box, Unlocked: Predictability and Understandability in Military AI*. Geneva, Switzerland: UNIDIR. https://unidir.org/publication/the-black-box-unlocked/.

Horowitz, Michael C. 2018. "Artificial Intelligence, International Competition, and the Balance of Power." *Texas National Security Review* 1 (3): 36–57. https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/.

Horowitz, Michael C. and Lauren Kahn. 2024. "Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts." *International Studies Quarterly* 68 (2): sqae020. https://doi.org/10.1093/isq/sqae020.

Horowitz, Michael C., Lauren Kahn and Casey Mahoney. 2020. "The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?" *Orbis* 64 (4): 528–43. https://doi.org/10.1016/j.orbis.2020.08.003.

Horowitz, Michael, Shira Pindyck and Casey Mahoney. 2024. "AI, the International Balance of Power, and National Security Strategy." In *The Oxford Handbook of AI Governance*, edited by Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young and BaoBao Zhang, 914–36. New York, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.55.

Hunter Christie, Edward, Amy Ertan, Laurynas Adomaitis and Matthias Klaus. 2023. "Regulating lethal autonomous weapon systems: exploring the challenges of explainability and traceability." *AI and Ethics* 4: 229–45. https://doi.org/10.1007/s43681-023-00261-0.

Johnson, James. 2024. "Finding AI Faces in the Moon and Armies in the Clouds: Anthropomorphising Artificial Intelligence in Military Human-Machine Interactions." *Global Society* 38 (1): 67–82. https://doi.org/10.1080/13600826.2023.2205444.

Klimburg, Alexander and Virgilio Almeida. 2019. "Cyber Peace and Cyber Stability: Taking the Norm Road to Stability." *IEEE Internet Computing* 23 (4): 61–6. https://doi.org/10.1109/MIC.2019.2926847.

Krasner, Stephen D. 1982. "Structural Causes and Regime Consequences: Regimes as Intervening Variables." *International Organization* 36 (2): 185–205.

Krieger, Miriam, Lynne M. Chandler Garcia, John H. Riley and Will Atkins, eds. 2021. *American Defense Policy*. Baltimore, MD: Johns Hopkins University Press.

Kwik, Jonathan and Tom Van Engers. 2021. "Algorithmic fog of war: When lack of transparency violates the law of armed conflict." *Journal of Future Robot Life* 2 (1–2): 43–66. https://doi.org/10.3233/FRL-200019.

La Fors, Karolina, Bart Custers and Esther Keymolen. 2019. "Reassessing values for emerging big data technologies: integrating design-based and application-based approaches." *Ethics and Information Technology* 21 (3): 209–26.

Lancaster, Caitlin M., Kelsea Schulenberg, Christopher Flathmann, Nathan J. McNeese and Guo Freeman. 2024. "'It's Everybody's Role to Speak Up... But Not Everyone Will': Understanding AI Professionals' Perceptions of Accountability for AI Bias Mitigation." *ACM Journal on Responsible Computing* 1 (1): 1–30. https://doi.org/10.1145/3632121.

Lewis, Dustin A. 2022. "On 'Responsible AI' in War: Exploring Preconditions for Respecting International Law in Armed Conflict." In *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*, edited by Silja Voeneky, Philipp Kellmeyer, Oliver Mueller and Wolfram Burgard, 488–506. Cambridge Law Handbooks. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/9781009207898.037.

Lewis, James Andrew. 2022. "Strengthening a Transnational Semiconductor Industry." Center for Strategic and International Studies, June 2. www.csis.org/analysis/strengthening-transnational-semiconductor-industry.

Linardatos, Pantelis, Vasilis Papastefanopoulos and Sotiris Kotsiantis. 2021. "Explainable AI: A Review of Machine Learning Interpretability Methods." *Entropy* 23 (1): 18–63. https://doi.org/10.3390/e23010018.

Lingevicius, Justinas. 2023. "Military artificial intelligence as power: consideration for European Union actorness." *Ethics and Information Technology* 25 (1): 18. https://doi.org/10.1007/s10676-023-09684-z.

Maas, Matthijs M. and José Jaime Villalobos. 2023. "International AI Institutions: A Literature Review of Models, Examples, and Proposals." SSRN. https://doi.org/10.2139/ssrn.4579773.

Mark, Jeremy and Dexter Tiff Roberts. 2023. "United States-China semiconductor standoff: A supply chain under stress." Atlantic Council, February 23. www.atlanticcouncil.org/in-depth-research-reports/issue-brief/united-states-china-semiconductor-standoff-a-supply-chain-under-stress/.

Mazarr, Michael J. 2022. "Understanding Competition: Great Power Rivalry in a Changing International Order — Concepts and Theories." RAND Corporation. March. www.rand.org/pubs/perspectives/PEA1404-1.html.

Meerveld, H. W., R. H. A. Lindelauf, E. O. Postma and M. Postma. 2023. "The irresponsibility of not using AI in the military." *Ethics and Information Technology* 25 (14): 1–6. https://doi.org/10.1007/s10676-023-09683-0.

Munn, Luke. 2023. "The uselessness of AI ethics." *AI and Ethics* 3: 869–77. https://doi.org/10.1007/s43681-022-00209-w.

Nadibaidze, Anna. 2023. "'Responsible AI' in the Military Domain: Implications for Regulation." *Opinio Juris* (blog), March 31. https://opiniojuris.org/2023/03/31/responsible-ai-in-the-military-domain-implications-for-regulation/.

Novelli, Claudio, Mariarosaria Taddeo and Luciano Floridi. 2023. "Accountability in artificial intelligence: what it is and how it works." AI & Society, February 7. https://doi.org/10.1007/s00146-023-01635-y.

Office of the High Commissioner on Human Rights. 2024. "Gaza: UN experts deplore use of purported AI to commit 'domicide' in Gaza, call for reparative approach to rebuilding." Press release, April 15. www.ohchr.org/en/press-releases/2024/04/gaza-un-experts-deplore-use-purported-ai-commit-domicide-gaza-call.

Oniani, David, Jordan Hilsman, Yifan Peng, Ronald K. Poropatich, Jeremy C. Pamplin, Gary L. Legault and Yanshan Wang. 2023. "Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare." *npj Digital Medicine* 6 (1): 225–35. https://doi.org/10.1038/s41746-023-00965-x.

Pacholska, Magdalena. 2023. "Military Artificial Intelligence and the Principle of Distinction: A State Responsibility Perspective." *Israel Law Review* 56 (1): 3–23. https://doi.org/10.1017/S0021223722000188.

Palmer, Alex W. 2023. "'An Act of War': Inside America's Silicon Blockade Against China." *The New York Times Magazine*, July 12. www.nytimes.com/2023/07/12/magazine/semiconductor-chips-us-china.html.

Persbo, Andreas. 2010. "The role of non-governmental organizations in the verification of international agreements." In *Disarmament Forum: Arms Control Verification*, edited by Kerstin Vignard and Jane Linekar, 65–73. Geneva, Switzerland: United Nations. https://unidir.org/publication/disarmament-forum-arms-control-verification/.

Redmon, Joseph, Santosh Divvala, Ross Girshick and Ali Farhadi. 2016. "You Only Look Once: Unified, Real-Time Object Detection." *arXiv*, May 9. https://doi.org/10.48550/arXiv.1506.02640.

Rommen, Rebecca. 2024. "Israel's 'Where's Daddy?' AI system helps target suspected Hamas militants when they're at home with their families, report says." Business Insider, April 7. www.businessinsider.com/israel-ai-system-wheres-daddy-strikes-hamas-family-homes-2024.

Russell, Stuart J. and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. Boston, MA: Pearson.

Russell, Stuart, Karine Perset and Marko Grobelnik. 2023. "Updates to the OECD's definition of an AI system explained." OECD.AI, November 29. https://oecd.ai/en/wonk/ai-system-definition-update.

Santoni de Sio, Filippo and Giulio Mecacci. 2021. "Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them." *Philosophy and Technology* 34 (4): 1057–84. https://doi.org/10.1007/s13347-021-00450-x.

Sassòli, Marco. 2002. "State responsibility for violations of international humanitarian law." *International Review of the Red Cross* 84 (846): 401–34. https://international-review.icrc.org/sites/default/files/S1560775500097753a.pdf.

Sastry, Girish, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe et al. 2024. "Computing Power and the Governance of Artificial Intelligence." *arXiv*, February 14. https://doi.org/10.48550/arXiv.2402.08797.

Scharre, Paul. 2023. *Four Battlegrounds: Power in the Age of Artificial Intelligence*. New York, NY: W. W. Norton & Company.

Scharre, Paul and Megan Lamberth. 2022. *Artificial Intelligence and Arms Control*. Washington, DC: Center for a New American Security. www.cnas.org/publications/reports/artificial-intelligence-and-arms-control.

Schaul, Kevin, Szu Yu Chen and Nitasha Tiku. 2023. "Inside the secret list of websites that make AI like ChatGPT sound smart." *The Washington Post*, April 19. www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

Schmitt, Lewin. 2022. "Mapping global AI governance: a nascent regime in a fragmented landscape." *AI and Ethics* 2: 303–14. https://doi.org/10.1007/s43681-021-00083-y.

Schreiner, Maximilian. 2023. "GPT-4 architecture, datasets, costs and more leaked." The Decoder, July 11. https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.

Sheikh, Haroon, Corien Prins and Erik Schrijvers. 2023. "Artificial Intelligence: Definition and Background." In *Mission AI: The New System Technology*, edited by Haroon Sheikh, Corien Prins and Erik Schrijvers, 15–41. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-031-21448-6_2.

Shivakumar, Sujai, Charles Wessner and Thomas Howell. 2023. "The Strategic Importance of Legacy Chips." Center for Strategic and International Studies, March 3. www.csis.org/analysis/strategic-importance-legacy-chips.

Soare, Simona R., Pavneet Singh and Meia Nouwens. 2023. "Software-defined Defence: Algorithms at War." International Institute for Strategic Studies, February 17. www.iiss.org/research-paper/2023/02/software-defined-defence/.

Stanley-Lockman, Zoe. 2021. "Military AI Cooperation Toolbox: Modernizing Defense Science and Technology Partnerships for the Digital Age." Center for Security and Emerging Technology Issue Brief. August. https://cset.georgetown.edu/publication/military-ai-cooperation-toolbox/.

Sugden, Robert. 1989. "Spontaneous Order." *Journal of Economic Perspectives* 3 (4): 85–97. www.jstor.org/stable/1942911.

Sweijs, Tim. 2023. "A Genealogy of Ultimata." In *The Use and Utility of Ultimata in Coercive Diplomacy*, edited by Tim Sweijs, 37–68. Cham, Switzerland: Springer International. https://doi.org/10.1007/978-3-031-21303-8_2.

Taddeo, Mariarosaria, David McNeish, Alexander Blanchard and Elizabeth Edgar. 2021. "Ethical Principles for Artificial Intelligence in National Defence." *Philosophy & Technology* 34: 1707–29.

Taddeo, Mariarosaria, Marta Ziosi, Andreas Tsamados, Luca Gilli and Shalini Kurapati. 2022. *Artificial Intelligence for National Security: The Predictability Problem*. Centre for Emerging Technology and Security Research Report. September. https://cetas.turing.ac.uk/publications/artificial-intelligence-national-security-predictability-problem.

Tallberg, Jonas, Eva Erman, Markus Furendal, Johannes Geith, Mark Klamberg and Magnus Lundgren. 2023. "The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research." *arXiv*, May 13. https://doi.org/10.48550/arXiv.2305.11528.

Thadani, Akhil and Gregory C. Allen. 2023. "Mapping the Semiconductor Supply Chain: The Critical Role of the Indo-Pacific Region." Center for Strategic and International Studies, May 30. www.csis.org/analysis/mapping-semiconductor-supply-chain-critical-role-indo-pacific-region.

The Hague Centre for Strategic Studies. 2018. "Artificial Intelligence and Its Future Impact on Security." Testimony prepared by Dr. Tim Sweijs for the Committee on Foreign Affairs and the Subcommittee on Security and Defense of the European Parliament, October 10. https://hcss.nl/wp-content/uploads/2021/01/HCSS_Security_Tim_Testimony_SEDECommittee_PDF-1-.pdf.

Tigard, Daniel W. 2021. "Responsible AI and moral responsibility: a common appreciation." *AI and Ethics* 1: 113–17. https://doi.org/10.1007/s43681-020-00009-0.

Trabucco, Lena and Matthijs M. Maas. 2023. "Technology Ties: the Rise and Roles of Military AI Strategic Partnerships." SSRN. https://doi.org/10.2139/ssrn.4629283.

Trusilo, Daniel. 2022. "Responsible AI Symposium – Implications of Emergent Behavior for Ethical AI Principles for Defense." Lieber Institute West Point, November 30. https://lieber.westpoint.edu/implications-emergent-behavior-ethical-artificial-intelligence-principles-defense/.

———. 2023. "Autonomous AI Systems in Conflict: Emergent Behavior and Its Impact on Predictability and Reliability." *Journal of Military Ethics* 22 (1): 2–17. https://doi.org/10.1080/15027570.2023.2213985.

UNODA. 2024. "Convention on Certain Conventional Weapons – Group of Governmental Experts on Lethal Autonomous Weapons Systems." https://meetings.unoda.org/ccw-/convention-on-certain-conventional-weapons-group-of-governmental-experts-on-lethal-autonomous-weapons-systems-2024.

van Hooft, Paul, Lotje Boswinkel and Tim Sweijs. 2022. *Shifting sands of strategic stability: Towards a new arms control agenda*. February. The Hague, The Netherlands: The Hague Centre for Strategic Studies.

Vestner, Tobias. 2022. "Responsible AI Symposium – The Nexus between Responsible Military AI and International Law." Lieber Institute West Point, November 17. https://lieber.westpoint.edu/nexus-between-responsible-military-ai-international-law/.

von Ingersleben-Seip, Nora. 2023. "Competition and cooperation in artificial intelligence standard setting: Explaining emergent patterns." *Review of Policy Research* 40 (5): 781–810. https://doi.org/10.1111/ropr.12538.

Walters, Robert and Marko Novak. 2021. *Cyber Security, Artificial Intelligence, Data Protection and the Law*. Singapore: Springer.

Wendt, Alexander. 1994. "Collective Identity Formation and the International State." *American Political Science Review* 88 (2): 384–96. https://doi.org/10.2307/2944711.

Winston, Carla. 2023. "International Norms as Emergent Properties of Complex Adaptive Systems." *International Studies Quarterly* 67 (3): 1–8. https://doi.org/10.1093/isq/sqad063.

Wolfrum, Rüdiger. 1987. "Reparation for Internationally Wrongful Acts." In *Encyclopedia of Disputes Installment 10*, edited by Rudolf L. Bindschedler, Thomas Buergenthal, Karl Doehring, Jochen Abr. Frowein, Günther Jaenicke, Herbert Miehsler, Hermann Mosler et al., 352–53. Amsterdam, The Netherlands: Elsevier. https://doi.org/10.1016/B978-0-444-86241-9.50089-0.

Zyberi, Gentian. 2018. "Enforcement of International Humanitarian Law." In *International Human Rights Institutions, Tribunals, and Courts*, edited by Gerd Oberleitner, 377–400. Singapore: Springer.

Annex

Table A1: Summary of Evaluations of Normative Focus Areas

Normative Focus Area	Strengths	Weaknesses	Evaluation
Accordance with international law	→ Intuitive starting point, which represents existing consensus → Foundation for other normative focus areas → Often open to interpretation, allowing for more actor buy-in	→ Lack of specificity as to which treaties or principles are relevant → Lack of clarity as to what extent existing international law is sufficient → Issues of interpretation of existing laws	**Adoptability:** High **Verifiability:** Low–medium **Enforceability:** Low
Responsibility and accountability	→ Reaffirms the relationship between responsibility and accountability → Fits within the human-centric international law framework → Recognizes that many actors are involved within the AI lifecycle	→ Lack of clarity as to where the responsibility of one actor ends and another actor begins → Lack of clarity as to whether AI should be governed within a framework of individual and state responsibility → (Only) negative formulation may discourage actors from taking on responsibility	**Adoptability:** High **Verifiability:** Medium **Enforceability:** Medium
Explainability and traceability	→ Recognizes the importance of understanding for responsibility and trust → Focuses attention on improving AI systems overall	→ Lack of clarity on how to account for the "double black-box" issue → Impossible to avoid more technical questions in political discussions → Lack of specificity about how much additional education and reskilling is needed	**Adoptability:** Medium **Verifiability:** Low (potentially high) **Enforceability:** Low (potentially high)
Bias and harm mitigation	→ Places safeguarding people at the core of all discussions → Clearly identifies and acknowledges certain types of harm and biases that pose a risk	→ Too broad as a category for issues to be addressed comprehensively → Focus on individuals rather than externalities → Focus on response rather than prevention → Some types of biases and harm remain unaccounted for	**Adoptability:** High **Verifiability:** High (if bias or harm prevention has failed) **Enforceability:** High (in accordance with international law)
Reliability	→ Explicitly recognizes the need for people to maintain ultimate say over AI systems → Outlines some responses to undesired outcomes of AI systems	→ Issues of subjective judgment (i.e., what is inappropriate behaviour?) → Lack of clarity as to how context and specific cases would be accounted for	**Adoptability:** High **Verifiability:** Medium (at state level) **Enforceability:** Medium (at state level)
Governability	→ Meta-agreement justifying the existence of governance initiatives → Foundation for further debate	→ Lack of clarity as to how national and international governance interact → Can be seen as an "empty" norm	**Adoptability:** Low **Verifiability:** Medium **Enforceability:** Low
Exchange of practices	→ Emphasizes value of innovation for AI → Proposes additional way of maintaining transparency	→ Difficult to adhere to in current geopolitical environment	**Adoptability:** Low **Verifiability:** Medium **Enforceability:** Low

Source: Authors.

Table A2: Overview of Attention Dedicated to Normative Focus Areas in Governance Initiatives Reviewed

Normative Focus Area	REAIM Summit Call to Action	Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy	(Draft) UNGA Resolution on Lethal Autonomous Weapons	Communiqué of the Latin American and the Caribbean Conference of Social and Humanitarian Impact of Autonomous Weapons	CARICOM Declaration on Autonomous Weapons Systems	(Summary) NATO Artificial Intelligence Strategy
Accordance with international law						
Responsibility and accountability						
Explainability and traceability						
Bias and harm mitigation						
Reliability						
Governability						
Exchange of practices						

Source: Authors.

Table A2: Continued

Normative Focus Area	Guiding Principles affirmed by the GGE on LAWS	(Draft) AUDA-NEPAD Artificial Intelligence Roadmap for Africa	OECD Recommendation of the Council on Artificial Intelligence	Bletchley Declaration	EU Artificial Intelligence Act	IEEE Position Statement on Ethical Aspects of Autonomous and Intelligent Systems	Charlevoix Common Vision for the Future of Artificial Intelligence
Accordance with international law							
Responsibility and accountability							
Explainability and traceability							
Bias and harm mitigation							
Reliability							
Governability							
Exchange of practices							

Centre for International
Governance Innovation

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org