Marijan, Branka

Working Paper

Through a glass, darkly: Transparency and military AI systems

CIGI Papers, No. 315

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

Suggested Citation: Marijan, Branka (2025) : Through a glass, darkly: Transparency and military AI systems, CIGI Papers, No. 315, Centre for International Governance Innovation (CIGI), Waterloo, ON, Canada

This Version is available at:
https://hdl.handle.net/10419/311807

Centre for International
Governance Innovation

CIGI Papers No. 315 — January 2025

Through a Glass, Darkly: Transparency and Military AI Systems

Branka Marijan

Centre for International
Governance Innovation

CIGI Papers No. 315 — January 2025

Through a Glass, Darkly: Transparency and Military AI Systems

Branka Marijan

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.

Credits

Managing Director and General Counsel Aaron Shull
Director, Program Management Dianna English
Program Manager and Research Associate Kailee Hilt
Senior Publications Editor Jennifer Goyder
Publications Editor Christine Robertson
Graphic Designer Sepideh Shomali

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

Table of Contents

About the Author

Branka Marijan is a CIGI senior fellow and a senior researcher at Project Ploughshares. She is a lecturer in the Master of Global Affairs program at the Munk School of Global Affairs and Public Policy at the University of Toronto.

At Ploughshares, Branka leads research on the military and security implications of emerging technologies. Her work examines concerns regarding the development of autonomous weapons systems and the impact of artificial intelligence and robotics on security provision. Her research interests include trends in warfare, civilian protection, use of drones and civil-military relations.

She holds a Ph.D. from the Balsillie School of International Affairs with a specialization in conflict and security. She has conducted research on post-conflict societies and published academic articles and reports on the impacts of conflict on civilians and diverse issues of security governance, including security sector reform.

Branka closely follows United Nations disarmament efforts and attends international and national consultations and conferences. She is a board member of the Peace and Conflict Studies Association of Canada and a research fellow at the Kindred Credit Union Centre for Peace Advancement at the University of Waterloo.

Executive Summary

International governance discussions on military artificial intelligence (AI) systems often emphasize the need for transparency. However, transparency is a complex and multi-faceted concept, understood in various ways within international debates and literature on the responsible use of AI. It encompasses dimensions such as explainability, interpretability, understandability, predictability and reliability. The degree to which these aspects are reflected in state approaches to ensuring transparent and accountable systems remains unclear and requires further investigation. Additionally, achieving transparency in military AI applications presents several challenges. First, the inherent opacity of the technology can make it difficult to trace and understand decision-making processes. Second, military institutions are more likely to adopt voluntary transparency measures that focus on ensuring operators have a general understanding of system functionality, without fully addressing the nuances of accountability. Furthermore, disparities in technological capabilities among states suggest uneven testing and training standards, complicating the evaluation of human decision making and accountability. Lastly, given the sensitivity of national defence and international security, military AI systems are expected to remain highly classified, making external evaluation difficult. This paper proposes pathways to overcome these challenges and outlines a framework for comprehensive transparency, which is essential for the responsible use of AI in military contexts.

Introduction

In international discussions on responsible military use of AI, transparency is frequently emphasized. Transparency is also a central concern across AI ethical principles in civilian contexts (Jobin, Ienca and Vayena 2019). However, the conceptualization of transparency varies considerably. For some governments, transparency entails some disclosure of information regarding the testing, evaluation and functioning of various systems by states. For others, it means that military AI systems must be sufficiently transparent to their own militaries and ensure that commanders understand their operations and can intervene when these systems produce errors or unpredictable outputs. In this way, the understanding of transparency is generally one of "the understandability and predictability of systems" (Endsley, Bolte and Jones 2003, 146; National Academies of Sciences, Engineering, and Medicine 2022). However, the challenge remains that these varying interpretations of transparency will become even more significant as states begin operationalizing responsible AI principles. These principles will be especially important for ensuring the responsible use of AI and autonomous systems by military forces.

Already in practice in contemporary conflict zones such as Ukraine and Gaza, commitments to having military commanders understand AI systems are being challenged due to the nature of the technology, the use of off-the-shelf technologies and the lack of clear guidelines regarding the extent to which such understanding is required. There is also a broader lack of disclosure about the types and sophistication of AI-enabled systems being used and how they function. Notably, the AI target generation and decision support systems used by the Israel Defense Forces (IDF) in Gaza have raised concerns as investigative reports publicized their use, leading to more questions about their function (Abraham 2024; Davies, McKernan and Sabbagh 2023). However, little information has been provided by Israel on how the systems function and the country has argued that it is not using AI systems to autonomously select targets without human involvement (Varella and Acheson 2024, 5). These assurances have not been seen as sufficient by those alarmed at the reports regarding Israeli systems. Transparency regarding AI and autonomous systems in the military domain also involves some ability to access information on systems, ideally to have these systems be evaluated

or audited, preferably by a reputable third party. Such an extensive evaluation and auditing, while likely to be done internally, is unlikely possible to be performed externally. As such, information sharing and confidence-building measures at the global level will need to be creatively developed.

Several questions arise when seeking to establish a deeper understanding of transparency that satisfies both international governance bodies and technical and operational requirements. Does the military commander need to understand how each node of the AI system is connected? Would a deep enough understanding be possible or required, and for which uses? What would be a sufficient level of understanding by the human operator or war fighter to ensure their clear accountability for actions aided by or carried out by an AI system? Additionally, what information needs to be shared among various governments to ensure confidence in the responsible use of AI and autonomous systems?

These questions are considerably more relevant as militaries are increasingly using AI systems across a variety of functions, including recruitment, training, logistics, equipment maintenance, surveillance and targeting (Grand-Clément 2023). The different uses will have varying requirements of transparency that serve different functions and satisfy ethical and legal requirements at various levels of governance. For some uses, like those described as "back-end" office functions such as recruitment, the requirements will primarily focus on ensuring fairness and privacy, as well as meeting various domestic laws on employing individuals (Taddeo et al. 2019). On the other end of the spectrum, and of most concern to this paper, are high-risk applications, such as the use of AI systems in decision support related to deployment of force or in weapon systems, with varying degrees of autonomy. The requirements will be more stringent, needing to meet internal and national standards as well as international legal requirements and governance mechanisms. The latter issue, while particularly critical to international security, remains the most challenging to address due to inherent security considerations.

Transparency in the military use of AI and autonomy at the global level faces several key obstacles. First, the inherent complexity of the technology, especially as systems become more advanced, learn and evolve, makes ensuring their understandability challenging in practice. There is an active debate on the extent to which systems need to be explainable as well as interpretable by humans and what degree of understanding is required by those deploying systems. Additionally, the dual-use nature of AI and the use of commercial off-the-shelf technologies and tools, as utilized in Ukraine, may introduce systems that have not been adequately tested for defence contexts. Second, while militaries are more inclined to commit to transparency measures that ensure operators understand the systems, broader transparency or allowing external evaluation of these systems remains significantly more challenging. Third, and relatedly, military AI systems are often closely guarded due to national security concerns. This confidentiality can hinder the willingness of states to share information regarding the capabilities of various systems. This tendency is particularly true with more adversarial nations, as transparency regarding the functioning of military AI systems is unlikely to be shared due to fears of exposing confidential technologies that may provide a technical edge to other state actors. Transparency, therefore, often collides with national security (Etzioni 2018).

This paper explores the feasibility of achieving transparency in military AI systems, identifies the associated challenges and proposes pathways to develop effective transparency mechanisms. It begins by examining differing definitions of transparency, from technical understandings to international security governance. It then discusses how these various approaches have emerged in the discourse on military AI governance. Drawing on these diverse perspectives, the paper proposes elements of a comprehensive transparency approach to consider for international governance mechanisms. Ultimately, transparency mechanisms in the most concerning military AI applications, such as decision making related to the use of force, will also require a layered set of governance commitments and confidence-building measures. These should include clear legally binding commitments, voluntary measures and exchanges of information. Finally, many military applications of AI are likely to remain shrouded in secrecy. However, achieving a satisfactory level of translucency in applications with the most significant impact on global security will greatly enhance global stability.

Transparency and Technical Understandings

Transparency as a concept is ever evolving and multi-dimensional (Ball 2009; Hansen, Christensen and Flyverbom 2015). Adding to the confusion are the differing ways that transparency is understood in technical literature on AI, most notably explainable AI research, and in discussions relating to the governance of AI. AI transparency in literature in the fields of science and engineering encompasses more technical views, including "algorithmic transparency" as well as the interpretability and explainability of models (Larsson and Heintz 2020). The two concepts as explained below are not synonymous. In more technical discussions, transparency is seen as a goal to ensure the proper functioning of systems, by being able to look into the internal functioning of systems and providing the ability to test and improve on the models and algorithms. It is also a way to show that the technology is trustworthy, leading to its wider adoption. There is also a slight differentiation between discussions of interpretability and explainability in more technical AI literature. Interpretability focuses on ensuring that humans operating the systems understand how the models function and the reasoning behind certain decisions (Petch, Di and Nelson 2022, 205).

Explainability involves developing a sufficient technical understanding of how a system works and arrives at decisions. This approach does not aim to capture every calculation made by the model but seeks to provide a clear and comprehensible explanation of its functioning (ibid.). The aim is to improve and recalibrate models as well as to be able to ensure accountability. However, with the use of deep neural networks, which are a result of highly complex non-linear statistical models, humans cannot understand how the system arrived at a given output, known as the "black-box" problem (Schraagen 2023, 1720; Hassija et al. 2024). Explainable AI aims to resolve the black-box problem and to find technical solutions to ensure the predictions or outputs often aiding highly critical decisions, such as the use of AI in military drones, are explainable enough to satisfy legal and operator requirements and interpretable by humans who would be fielding these types of systems. Explainability and interpretability are combined in efforts to develop transparent "glass-box" AI systems, which enhance the trustworthiness of these technologies and promote more ethical applications (Franzoni 2023). However, despite these efforts, explainable AI does not offer a solution to greater transparency into how different models function and often involves trade-offs in prioritizing certain aspects for greater clarity.

Given the concerns about the lack of understanding of most deep neural networks, there is a broader question about justification for use of these models in safety-critical applications in the first place (Hassija et al. 2024, 47). In the military domain, it is one that states will ultimately need to decide and, ideally, develop international regulations to address. Still, a significant number of research projects are devoted precisely to technical solutions for black-box models. For example, surrogate models can help approximate an output of a black-box model "by developing a new interpretable model, such as a logistic regression or a short decision tree, on the predictions of the black-box model" (Petch, Di and Nelson 2022, 207). The limitation with this method, as well as others, is that these are approximations and with many complex layers of decision making involved, they may be a snapshot of what is actually happening "inside" the model (ibid., 209). As a result, some scholars suggest that what much of the explanatory AI research offers is an "ersatz understanding" of the black-box models (Babic et al. 2021).

Mike Ananny and Kate Crawford (2018, 974) also caution against simplified technical understandings of transparency, highlighting the limitations to its idealized understandings. They point out that, "rather than privileging a type of accountability that needs to look inside systems, that we instead hold systems accountable by looking *across* them — seeing them as sociotechnical systems that do not *contain* complexity but *enact* complexity by connecting to and intertwining with assemblages of humans and non-humans." In their view, to avoid narrow definitions that allude to understanding what systems are doing, there needs to be a broader consideration of the limitations of such understandings of transparency. Acknowledging the limitations can then lead to decisions regarding the design and use of systems that, in fact, lead to greater accountability. For example, they argue that "if transparency assumes the active participation of individuals interested in and able to provide oversight, then the model of accountability might ask whether they have the motivations, skills, and

associations required to achieve the collective oversight transparency promises" (ibid., 984). This consideration is particularly relevant for the use of military AI systems, as requirements for human oversight or control can be more clearly detailed to ensure transparency is implemented in practice. Securing this oversight would involve additional best practice exchanges and the establishment of norms requiring the training of individuals deploying particular systems.

Transparency and International Security Governance

In international governance and arms control discussions, transparency is largely associated with the availability of, and access to, certain information and involves mechanisms for sharing that information. According to Björn Hagelin et al. (2006), transparency is about the release of information by those who possess it, notably governments. Bernard I. Finel and Kristin M. Lord (2000, 3) note that "transparency in the political realm is a condition in which information about governmental preferences, intention and capabilities is made available either to the public or other outsiders." In international security, and particularly related to weapons, such transparency has often been fraught, but it has existed in various arms control efforts. For example, information regarding arms transfers is disclosed, albeit inconsistently (Holtom, Mensah and Nicolin 2022). Even regarding nuclear weapons, some states have exchanged information or disclosed the number of nuclear weapon delivery systems and the approximate number of available nuclear warheads (Grand 2003, 42). The disclosure is seen as a confidence-building measure, which helps to establish trust among states as information becomes shared, providing a level of knowledge about the weapons available to particular states.

Providing information on the conventional arms trade and nuclear capabilities is naturally highly sensitive, as more advanced militaries are unwilling to share details that could reveal the functioning or vulnerabilities of their systems. This concern is also prominent with emerging military technologies such as AI, given the belief in AI's centrality to future state power (Scharre 2023a). Placing AI at the centre of the future geopolitical order has meant that most states are reluctant to support the development of strong, legally binding instruments requiring comprehensive testing and evaluation mechanisms or to even share information regarding developments of AI-enabled systems. The increased competition between the great powers, notably the United States and China, has also placed new technologies and AI at the centre of the rivalry.

As a result, transparency in military AI applications is often understood narrowly, focusing on what is knowable about how systems function by the military's own operators or war fighters, termed "internal transparency." States might share information about different systems between various relevant government departments or develop transparency mechanisms among a close group of allies within various security arrangements. However, the willingness for this broader openness, even among allies, or "external transparency," remains unclear. Nonetheless, there is increased recognition that the sharing of information could be an important confidence-building measure, for example, in relation to the testing and evaluation of military AI systems (Horowitz and Scharre 2021). Even so, in testing and evaluation, countries express concern about the level of detail that is shared.

Too much openness about how systems function can also be dangerous, as it might reveal pathways for malicious actors to poison data or interfere with system operations. Political considerations also play a role in achieving transparency in defence contexts, with democratic societies being more forthcoming with information than authoritarian regimes (Grand 2003). Even among democratic societies, there is variance in the quality and reliability of the information provided. Despite its limitations, ths information enables states to make better decisions and transparency has become a foreign policy instrument (Yordanova 2015).

As such, despite the growing great power competition, there are possible strategic considerations for engaging with transparency mechanisms. James Marquardt (2011) notes that in international military-security affairs, powerful countries such as the United States engage in transparency efforts for their own

strategic purposes. Marquardt (ibid., 11) argues that in international relations, power politics largely shape the discussions about the demands for increased transparency and the strategies countries implement to achieve it. Simply put, the extent and manner in which states provide transparency is influenced by their strategic interests and positions on the global stage. Tsvetelina Yordanova (2015, 4) summarizes some of the reasons why more powerful states sustain transparency efforts, such as entangling non-democratic countries into arrangements requiring information exchanges, exerting soft power, influencing the domestic policies of other countries, and opening opportunities for international trade and investment.

However, Yordanova also notes that developing countries benefit from transparency efforts. These benefits include demonstrating good faith to gain support for further domestic development and investment initiatives, building stronger institutions, improving chances for external financial support and enhancing their ability to play a role in broader governance mechanisms (ibid.). Beyond these considerations, transparency efforts among various countries can contribute to greater regional stability in more contested areas. The greater exchange of information can also enable civil society organizations to more effectively monitor adherence to international agreements.

Military AI Governance Discussions

Considering the potential benefits of building confidence and reducing tensions through greater information exchanges, transparency has been recognized as a principle of responsible military AI use in various international discussions and national defence commitments. Perhaps the most prominent example of calls for internal transparency is the US-led Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, released in 2023 and endorsed by 54 states thus far. The focus on building transparent systems in the declaration is then interpreted in a narrower sense. Under measure F, the declaration says, "states should ensure that military AI capabilities are developed with methodologies, data sources, design procedures, and documentation that are transparent to and auditable by their relevant defense personnel" (Vergun 2023). The emphasis is on ensuring that defence personnel can understand, document and audit these systems. In this way, the declaration acknowledges that different states will develop their own internal procedures for transparency, leading to varying standards based on technical capabilities. Hence, there needs to be some coordination and guidance to ensure compatibility.

In this vein, several working groups have been established as part of the political declaration effort to develop best practices and build capacity. Canada, alongside Portugal, has been tasked with leading the working group on accountability and transparency. This effort aims to contribute to the development of a process of operationally spelling out the practices among like-minded states. While this is an important piece of the governance framework on responsible military AI, other governance efforts will need to include a wider range of states beyond the endorsing states that are largely like-minded. The working group could potentially share its recommendations with states beyond those that have signed onto the declaration.

Transparency has also been highlighted at the United Nations Convention on Certain Conventional Weapons (CCW) meetings on lethal autonomous weapons systems (LAWS). At the March 2024 meeting of the Group of Governmental Experts of the CCW on LAWS, the United States noted that the CCW is a "uniquely valuable forum to promote transparency and the international discussion" on autonomous weapons (Acheson and Varella 2024, 29). Here the United States appears to be using a broader notion of transparency in the more political sense. Indeed, this role for the CCW had been noted by various states and civil society.

In 2015, the Stockholm International Peace Research Institute (SIPRI) pointed out that it was not too early for the CCW to consider both how to share information as well as what information should be shared (SIPRI 2015). SIPRI also noted that greater transparency measures could be focused on the development and acquisition of autonomous weapon systems as well as their use. To move the discussion at the CCW forward, SIPRI proposed that states exchange information on, for example, the existing standards and procedures for deploying weapons and with the degree and type of human control present

in various systems (ibid., 4). As such, this would involve states providing some insights into internal transparency mechanisms; sharing it in the CCW would provide a degree of external transparency.

Transparent AI for the Defence Context

Aside from the diplomatic discussions, for most militaries, transparency in how a system is functioning is meant to ensure reliability of systems as well as to address any unanticipated actions. For militaries, a key focus is having humans who understand what the systems are doing and also systems that can be controllable to a satisfactory degree. But what is the extent of the understanding required by humans regarding how the system is functioning? Do the human operators need to be able to explain each action? Addressing these questions is both about ensuring reliability of systems as well as assigning accountability.

According to the US National Academies of Sciences, Engineering, and Medicine (2022), transparency in the military domain involves two key components: display transparency and explainability. They note that "Display transparency: Provides a real-time understanding of the actions of the AI system as a part of situation awareness (SA). Explainability: Provides information in a backward-looking manner on the logic, process, factors, or reasoning upon which the system's actions or recommendations are based" (ibid., 36). There needs to be clarity both in terms of which tasks are delegated to AI systems as well as a sufficient understanding and predictability in how the system is functioning (Endsley 2023).

In terms of which tasks are assigned to AI, humans using and deploying AI systems related to high-risk applications, such as selection and engagement of targets, are apt to be required to understand in real time what the actions are that the system is performing and also be able to know how the system will react to some degree. Technical components such as interfaces need to be considered when using different systems. Display transparency is then a critical component of any definition of transparency that needs to be developed in international regulation on military applications of AI. Namely, it should be a required component for any decision making related to deployment of force, such as decision making regarding targeting, to ensure that it is humans who are sufficiently in control of these critical processes. This transparency requirement will then need to supplement a legally binding requirement for human control over critical processes. The National Academies of Sciences, Engineering, and Medicine acknowledges that display transparency will be most relevant and necessary to time-constrained applications, while explainability is most likely when there is sufficient time to examine the systems (ibid., 36).

Explainability, Understandability and Predictability

During international discussions, states and militaries have often made reference to the need for explainable systems. However, the usefulness of explainability, as it is understood in the scientific literature discussed earlier, has been debated in the military context. Nathan Gabriel Wood (2024) argues that explainability is perhaps less useful to those actually deploying the systems than to the designers and troubleshooters.

Instead, he suggests that the focus in deployment of AI in military contexts should be on human-machine teaming rather than explainability. Wood (ibid.) notes that "explainability may undermine efforts to improve human-machine teamings by creating a *prima facie* sense that the AI, due to its explainability, may be utilized with little (or less) potential for mistakes." His concern is primarily that the focus on explainable systems overlooks the need to ensure that those deploying the systems are adequately trained and interact appropriately with the AI tools. Still, even in human-machine teaming, it is essential for humans to understand the underlying factors, such as the basis for the system's reported accuracy rates.

In Wood's view, the crucial concern is predictability. The military commander does not necessarily need to be able to explain how a system functions: knowing "when it will function correctly and when it won't, and responding accordingly to

that knowledge, will suffice for the ethical and legal use of these systems" (ibid.). Here, Wood simply highlights that the soldiers deploying AI-enabled systems do not need to be AI developers or programmers. This view is also expressed by some military commanders. Consider, for example, an IDF colonel who noted that it can be hard to know how certain decisions were made by an AI-enabled system he was using. He stated, "And then sometimes I'm willing to say I'm satisfied with traceability, not explainability. That is, I want to understand what is critical for me to understand about the process and monitor it, even if I don't understand what every 'neuron' is doing" (Newman 2023). However, for *ex post* accountability, some degree of explainability is needed to establish whether it was truly the system that malfunctioned or if the human was responsible for particular decisions that perhaps led to a potential war crime. For accountability to be established, the soldiers do need to be aware and adequately trained on systems they are using.

For Arthur Holland Michel (2020), understandability and predictability are also crucial. Rather than explainability of systems, he notes that a better term to use is understandability. In his view, the "broader and more neutral term 'understandability' covers the technical explainability and interpretability of the AI system while also accounting for the human subject's capacity for understanding — and it does not imply agency on the part of the machine" (ibid., 1). It is not enough then to have technical explanations for how a system arrived at a decision, but rather that those deploying the system can also understand the actions of the system without a significant level of technical expertise. Understandability can at times be quite broad as military commanders currently are unlikely to understand how each weapons system performs certain actions to a granular degree, but they can generally understand how critical actions are performed.

Most importantly, military commanders currently can generally know how the system will perform in various conditions. Predictability is then a related component of understandability as it ensures that the outputs and effects of actions can be expected and anticipated (ibid., 5). Holland Michel notes that the technical definition includes expectations that the system will act in ways reflective of the testing or the training data (ibid.). In the operational sense, it is important for those using autonomous systems to be able to anticipate different actions. Holland Michel acknowledges the challenges of truly anticipating every scenario, particularly with more complex systems. Hence, to truly have responsible applications of military AI, Holland Michel argues that systems need to be highly understandable and predictable. In intense, high-pressure environments such as war zones, militaries are likely to want simplified explanations and the appeal of a technological edge may lead to the deployment of less understandable and predictable systems. However, such a scenario needs to be avoided, and states need to consider ways to ensure both understandability and predictability in the systems that they are planning on deploying at various stages of deployment and development.

Transparency and Operational Challenges

Still, achieving transparency in military AI systems is easier said than done. Transparency in the military AI domain can exist at different points in time and, conversely, become murkier at others. Simply put, it is not a constant condition. It exists on a spectrum and fluctuates depending on various factors such as the stage of development, the conditions of deployment and the operational context. For example, while those initially developing and testing AI-enabled systems for defence contexts can ensure a degree of knowing how the systems work, these systems can then be deployed in settings that vastly differ from the training conditions. An example would be taking AI-enabled reconnaissance systems currently being used in Ukraine and deploying them in desert conditions. While the system may have performed well and been tested rigorously in the Ukrainian context, the deployment in an environment quite different from the one where it was tested and deployed will raise concerns about predictability in functioning.

This scenario underscores the necessity for adaptive testing and re-certification processes. According to some experts, there should be stringent requirements for constraining technology to the specific domains for which it has been tested (Cummings 2024). If a system is to be used

outside of its original context, it should undergo a certification or approval process to ensure that it can function reliably under new conditions. This process would help maintain a higher level of understandability and predictability, mitigating the risks associated with deploying AI systems in unfamiliar environments. By limiting the deployment of such systems to their tested domains, military planners can avoid the pitfalls of unexpected system behaviour in uncharted environments. This approach would necessitate ongoing dialogue between developers, testers and end-users to continuously assess and update the operational parameters of AI systems. This certification process is likely to be developed internally by various states, and sharing information regarding how these states are carrying out the certification process is not to be expected.

Transparency in military AI also demands continuous monitoring and evaluation. Once an AI system is deployed, it should be subject to ongoing scrutiny to detect and address any deviations from its expected behaviour. This monitoring should involve collecting data on the system's performance, analyzing its decisions and making any necessary adjustments. Continuous monitoring helps maintain a high level of transparency and allows for prompt corrective actions when issues arise. In practice, this could mean setting up dedicated teams to oversee the operation of AI systems in the field. These teams would be responsible for tracking performance metrics, investigating anomalies and ensuring compliance with established standards. By maintaining a feedback loop between developers and operators, the military can ensure that AI systems remain transparent and reliable throughout their operational life cycle.

Despite these efforts, existing and, in all likelihood, future AI systems, are bound to struggle with the defence context, which is both generally unpredictable and also involves the reality of adversarial states seeking to infiltrate and undermine AI-enabled systems being deployed. Jon R. Lindsay (2023/2024, 45) notes that "adversaries have incentives to move conflict in unexpected directions, i.e., where AI systems have not been trained and will likely perform in undesired or suboptimal ways. This creates not only data problems but judgment problems as well." As such, in practice, maintaining the initial conditions of testing and evaluating systems is going to be difficult. It will likely lead to further pressure on human-machine teams as humans might not have been trained to anticipate outputs that systems will make in differing conditions. Nonetheless, as militaries continue to field systems from loitering munitions with computer vision to AI-piloted fighter jets, effort needs to be put toward achieving a semblance of transparency. This transparency then needs to compliment regulations and other governance mechanisms

Toward a Comprehensive Transparency Approach for Military AI

Technical and Internal Requirements

For internal transparency, militaries need to consider the life cycle of military systems and the stages from research and development (R&D) to testing and then deployment. Each stage, depending on application, will require specific requirements of transparency that may serve several purposes for internal military compliance and also, depending on the type of transparency, may provide assurance for other states as well.

Research, Development and Testing

The R&D of emerging technologies in the military domain is conducted behind closed doors. However, defence departments can internally ensure that systems undergoing R&D are adequately documenting data inputs, considering various environments and addressing adversarial strategies that could be deployed against these systems, among other considerations. There are also various systems and types of tasks or functions that these systems will need to perform using different AI methodologies (Holland Michel 2020, 6). Systems with a higher likelihood of unpredictable behaviour need closer examination, and considerations must be made regarding the international legal and normative requirements for some of these systems.

Given that AI is a general-purpose technology, the transfers of technology from the civilian to the military domain is probable (Scharre 2023b). As

such, rigorous testing of civilian AI technologies before they are deployed in military settings will be crucial. A generally acknowledged aspect of AI systems is their brittleness; while the systems can perform well in narrow applications, they struggle with more general applications (Mayer 2023). Thus, civilian systems might serve as the basis for military applications, but the transition from civilian to military use is fraught with potential pitfalls. Without thorough testing, these systems may exhibit unforeseen vulnerabilities or biases that could have catastrophic consequences in a combat environment. These consequences include the targeting ofcivilians and civilian infrastructure.

In civilian contexts, AI systems are subject to extensive testing and validation to ensure that they perform as intended. This process involves evaluating the system's accuracy, reliability and resilience under various conditions. When these systems are adapted for military use, they must undergo an even more stringent testing regime. The stakes are higher in military applications, where the margin for error is minimal and the potential for harm is significant.

Proper testing should encompass not only technical performance but also ethical considerations. For instance, the biases inherent in many AI systems can be amplified in a military context, leading to unjust outcomes. Therefore, testing protocols must include assessments of fairness and bias mitigation strategies.

Comprehensive Training for Users

Transparency in military AI systems also depends on the comprehensive training of individuals who operate these technologies. It is not enough for operators to merely understand how to use the systems; they must also grasp the underlying principles guiding the AI's decision-making processes (Lyons et al. 2017). The operators then need to share their knowledge regarding how the systems function with military commanders. This knowledge is crucial for ensuring that human commanders remain in control, can make informed decisions when interfacing with AI and have a strong degree of understandability of the AI systems.

Training programs should be designed to provide operators and commanders with a deep understanding of the AI systems that they are using. This training should include not only the systems' technical aspects but also the ways in which they may be designed to influence or "nudge" decision making (Millar 2015). For instance, AI systems often present data in ways that can subtly guide operators toward specific conclusions or actions. Recognizing these nudges is essential for maintaining human oversight and preventing the overreliance on AI recommendations.

Moreover, training should emphasize the importance of critical thinking and ethical considerations. Operators must be able to question and evaluate the AI's outputs, rather than accepting them at face value. As such, military organizations — where the traditional hierarchical structure can sometimes discourage questioning of automated systems — must undergo a cultural shift. By fostering a culture of critical engagement, the military can ensure that AI systems are used responsibly and transparently.

Continuous Assignment of Accountability

As military AI systems evolve and are updated, the assignment of accountability must be a continuous and dynamic process. One of the significant challenges in AI governance is ensuring that accountability is maintained throughout the life cycle of the system, including not only the initial deployment, but also subsequent updates and modifications.

When AI systems are updated, new features or adjustments can introduce unforeseen risks or alter the system's behaviour in ways that are not immediately apparent. Therefore, it is essential to have mechanisms in place for re-evaluating the system's performance and ethical implications after each update. This process should involve a diverse group of stakeholders, including technical experts, ethicists and end-users.

Moreover, clear lines of accountability must be established for every stage of the AI system's life cycle. This accountability includes identifying who is responsible for designing, testing, deploying and maintaining the system. In cases where the AI system makes a critical error or exhibits unintended behaviour, there should be a transparent process for determining responsibility and implementing corrective measures. As noted earlier, systems need to have a degree of explainability to ensure that those looking at *ex post* accountability can show whether a system malfunctioned or whether

the human decision maker acted in a manner that led to a particular crime being committed.

The assignment of accountability should also extend to the procurement process. Military organizations must ensure that contractors and vendors adhere to stringent ethical and transparency standards. The process should include requiring detailed documentation of the AI system's development process, testing protocols and any known limitations or biases. By holding vendors accountable, the military can enhance the overall transparency and reliability of its AI systems. Militaries will also need to develop procedures to evaluate the self-certification that vendors are likely to present and propose, as these can be manipulated or "gamed" by the developers (Say 2024).

External Transparency and Governance Frameworks

Internal transparency, while necessary, is not sufficient to address the legal and ethical challenges posed by the deployment of AI-enabled systems by militaries. To ensure responsible use, global standards and opportunities for confidence-building measures are also needed. External transparency can involve a diverse group of stakeholders. When states disclose information regarding their policies and the technical aspects of AI systems, the global scientific community can weigh in with insights from other safety-critical fields. This collaborative approach allows for a comprehensive evaluation of the issues, ensuring that multiple perspectives are considered in addressing the challenges of AI in the military domain.

Regulation must guide transparency mechanisms, with states deciding which systems and processes are acceptable to be enabled or carried out by AI technologies. At the international level, a legally binding instrument providing clear red lines will be necessary. Such a process could happen within the framework of the United Nations, which has fora representative of the widest number of countries. Additionally, there are established fora where states can exchange information and provide insights into their approaches and policies on the applications of AI in the military domain. The CCW is one such forum likely to continue dialogue on autonomous weapons. Beyond the CCW, the US-led political declaration and the working group on accountability and transparency can contribute to developing best practices and templates for states to use in tracking how understandability and predictability can be maintained through various stages of development.

Summits and international meetings can also help establish norms on the types of information that can be shared. For example, the Summit on Responsible AI in the Military Domain (REAIM) held in The Hague, Netherlands in 2023 and in Seoul, South Korea in 2024, brings together a broader group of states, including China and multiple stakeholders. REAIM is thus well positioned to provide a space for dialogue without pressure for specific regulations or more concrete commitments, and it features a multi-stakeholder environment (Csernatoni 2024). More exchanges between governments and in military-to-military dialogues can be helpful in understanding which governments are technically capable, but have received less attention regarding their positions in what they view as the responsible use of military AI, for example India. Although India has spoken about the responsible use of AI, it also signalled its wish to deploy these technologies more widely in CCW discussions. Indeed, greater clarity on military AI policies can contribute to knowledge and trust building among states.

Confidence-building measures are therefore critical to external transparency. Michael C. Horowitz, Lauren Kahn and Casey Mahoney (2020) examine the way in which confidence-building measures have historically contributed to international security and arms control agreements. They point out that while old approaches are not exact templates for military AI, these approaches offer considerations necessary for international cooperation. They note that states can exchange a degree of technical information on various systems and their policies that will foster a greater willingness to engage in dialogue, even among more adversarial states. Sharing of information about ethical considerations in various fora can also then provide insights on where various states stand. Thus, confidence-building measures are going to become more important as external evaluation or certification by third-party

institutions does not appear to be possible in the near term. Ioana Puscas (2022) points out that risk-based approaches can provide some focus on prioritizing confidence-building efforts in military applications of AI. Puscas provides the example of states agreeing to constraints on deploying AI in areas where the risk is exceptionally high, such as nuclear weapons. Such an agreement could perhaps lead to a risk-based governance framework for autonomous weapons and military AI, which could ensure that systems posing unacceptable risks are either prohibited or subjected to strict restrictions on their use (Marijan 2021).

Conclusion

The need for transparency in military AI systems extends beyond immediate operational concerns. It is also a crucial factor in maintaining public trust and international stability. As AI technologies become ever more integrated into military operations, the potential for misuse or unintended consequences increases. Due to the hype surrounding the technology and geopolitical realities, there might be pressure to deploy technologies that are not appropriately tested or understood. Transparent practices can help mitigate these risks by ensuring that AI systems are used ethically and responsibly. Transparency fosters confidence that these technologies are being used in accordance with ethical standards and that there are robust mechanisms in place to address any issues that arise. This trust is particularly important in democratic societies where public opinion can influence defence policies.

On the international stage, transparency in military AI systems can help prevent escalation and reduce the risk of conflict. When states are open about their AI capabilities and the measures that they have in place to ensure ethical use, it can build mutual trust and facilitate cooperation. Conversely, a lack of transparency can lead to suspicion and arms races, as countries may feel compelled to develop their AI capabilities in secret to maintain a strategic advantage.

Finally, transparency efforts can only supplement, not replace, regulation. The use of AI-enabled systems in the deployment of force is too consequential for global security to forgo legally binding instruments and other measures at the international level. Technical solutions have their limitations, and AI-enabled systems are likely to encounter significant operational issues outside of laboratory settings. States must consider the broader impacts of deploying these systems. As retired General Mark Milley, former chairman of the US Joint Chiefs of Staff, points out, “The idea that war is antiseptic and there are wonder weapons out there, that we can somehow make it painless…to think that technology’s going to resolve the horrors of war. It’s not” (quoted in Freedberg 2024). Hence, transparency efforts are one critical part of the broader governance and regulatory framework that needs to be developed for military applications of AI and autonomy. Without a comprehensive transparency framework, the potential for deployment of technologies that are not suitable for various environments or that increase risk of conflict escalation only rises. Contemporary conflicts already serve as a warning sign that assumptions about what is acceptable under existing norms and international laws are being eroded.

Works Cited

Abraham, Yuval. 2024. "'Lavender': The AI machine directing Israel's bombing spree in Gaza." *+972Magazine*, April 3. www.972mag.com/lavender-ai-israeli-army-gaza/.

Acheson, Ray and Laura Varella. 2024. "Topic 3: Risk Mitigation and Confidence Building." *CCW Report* 12 (1): 23–30. https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2024/gge/reports/CCWR12.1.pdf.

Ananny, Mike and Kate Crawford. 2018. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability." *New Media and Society* 20 (3): 973–89. https://doi.org/10.1177/1461444816676645.

Babic, Boris, Sara Gerke, Theodoros Evgeniou and I. Glenn Cohen. 2021. "Beware explanations from AI in health care." *Science* 373 (6552): 284–6. www.science.org/doi/10.1126/science.abg1834.

Ball, Carolyn. 2009. "What Is Transparency?" *Public Integrity* 11 (4): 293–308. https://doi.org/10.2753/PIN1099-9922110400.

Csernatoni, Raluca. 2024. "Governing Military AI Amid a Geopolitical Minefield." Carnegie Endowment Europe. July 17. https://carnegieendowment.org/research/2024/07/governing-military-ai-amid-a-geopolitical-minefield?lang=en&center=europe.

Cummings, Maisy. 2024. "How to ensure responsible use of AI in military decision-making – drawing lessons from the autonomous weapons systems (AWS) debate." REAIM Summit presentation. Seoul, South Korea.

Davies, Harry, Bethan McKernan and Dan Sabbagh. 2023. "'The Gospel': how Israel uses AI to select bombing targets in Gaza." *The Guardian*, December 1. www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets.

Endsley, R. Mica. 2023. "Supporting Human-AI Teams: Transparency, explainability, and situation awareness." *Computers in Human Behavior* 140, 107574. https://doi.org/10.1016/j.chb.2022.107574.

Endsley, R. Mica, Betty Bolte, and Debra G. Jones. 2003. *Designing for Situation Awareness: An Approach to Human-Centered Design*. London, UK: Taylor and Francis.

Etzioni, Amitai. 2018. "The Limits of Transparency." In *Transparency, Society and Subjectivity*, edited by Emmanual Alloa. and Dieter Thomä, 179–201. Cham, Switzerland: Palgrave Macmillan.

Finel, Bernard I. and Kristin M. Lord. 2000. *Power and Conflict in the Age of Transparency*. New York, NY: Palgrave Macmillan.

Franzoni, Valentina. 2023. "From Black Box to Glass Box: Advancing Transparency in Artificial Intelligence Systems for Ethical and Trustworthy AI." In *Computational Science and Its Applications – ICCSA 2023 Workshops*, edited by Osvaldo Gervasi, Beniamino Murgante, Anna Maria A. C. Rocha, Chiara Garau, Francesco Scorza, Yeliz Karaca and Carmelo M. Torre. Lecture Notes in Computer Science series, 14107: 118–30.

Freedberg, Sydney, Jr. 2024. "Can tech reduce civilian deaths in conflict? Mark Milley isn't so sure." Breaking Defense, May 8. https://breakingdefense.com/2024/05/mark-milley-civilian-death-ai-technology-alex-karp-palantir/.

Grand, Camille. 2003. "Nuclear Weapon States and the Security Dilemma." In *Transparency in Nuclear Warheads and Materials*, edited by Nicholas Zarimpas, 32–49. Oxford, UK: Oxford University Press and SIPRI.

Grand-Clément, Sarah. 2023. *Artificial Intelligence Beyond Weapons: Application and Impact of AI in the Military Domain*. Geneva, Switzerland: United Nations Institute for Disarmament Research. https://unidir.org/wp-content/uploads/2023/10/UNIDIR_AI_Beyond_Weapons_Application_Impact_AI_in_the_Military_Domain.pdf.

Hagelin, Björn, Mark Bromley, John Hart, Shannon N. Kile, Zdzislaw Lachowski, Wuyi Omitoogun, Catalina Perdomo, Eamon Surry and Siemon T. Wezeman. 2006. "Transparency in the arms life cycle." In *SIPRI Yearbook 2006: Armaments, Disarmament and International Security*, 245–67. Stockholm, Sweden: SIPRI. www.sipri.org/sites/default/files/YB06%20245%2006.pdf.

Hansen, Hans Krause, Lars Thøger Christensen and Mikkel Flyverbom. 2015. "Introduction: Logics of transparency in late modernity: Paradoxes, mediation and governance." *European Journal of Social Theory* 18 (2): 117–31. https://doi.org/10.1177/1368431014555256.

Hassija, Vikas, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud and Amir Hussain. 2024. "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence." *Cognitive Computation* 16: 45–74. https://doi.org/10.1007/s12559-023-10179-8.

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

Holtom, Paul, Anna Mensah and Ruben Nicolin. 2022. "The Case for Strengthening Transparency in Conventional Arms Transfers." Arms Control Today. November. www.armscontrol.org/act/2022-11/features/case-strengthening-transparency-conventional-arms-transfers.

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 and Paul Scharre. 2021. *AI and International Stability: Risks and Confidence-Building Measures.* Center for a New American Security. January. www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

Jobin, Anna, Marcello Ienca and Effy Vayena. 2019. "The global landscape of AI ethics guidelines." *Nature Machine Intelligence* 1: 389–99. https://doi.org/10.1038/s42256-019-0088-2.

Larsson, Stefan and Fredrik Heintz. 2020. "Transparency in artificial intelligence." *Internet Policy Review* 9 (2). https://policyreview.info/concepts/transparency-artificial-intelligence.

Lindsay, Jon R. 2023/2024. "War Is from Mars, AI Is from Venus: Rediscovering the Institutional Context of Military Automation." *Texas National Security Review* 7 (1): 29–47. https://doi.org/10.26153/tsw/50674.

Lyons, Joseph B., Matthew A. Clark, Alan R. Wagner and Matthew J. Schuelke. 2017. "Certifiable Trust in Autonomous Systems: Making the Intractable Tangible." *AI Magazine* 38 (3): 37–49. https://doi.org/10.1609/aimag.v38i3.2717.

Marijan, Branka. 2021. "Unacceptable risk and autonomous weapons." *The Ploughshares Monitor* 42 (4). www.ploughshares.ca/publications/unacceptable-risk-and-autonomous-weapons.

Marquardt, James J. 2011. *Transparency and American Primacy in World Politics.* Abingdon, UK: Ashgate.

Mayer, Michael. 2023. "Trusting machine intelligence: artificial intelligence and human-autonomy teaming in military operations." *Defense & Security Analysis* 39 (4): 521–38. https://doi.org/10.1080/14751798.2023.2264070.

Millar, Jason. 2015. "Expert testimony provided by Jason Millar to the Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWs), held within the framework of the UN Convention on Certain Conventional Weapons (CCW), Geneva, Switzerland, April 15, 2015." https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2015)/Jason%2BMillar%2B-%2BMeaningful%2BHuman%2BControl%2Band%2BDual-Use%2BTechnolo.gy.pdf.

National Academies of Sciences, Engineering, and Medicine. 2022. *Human-AI Teaming: State-of-the-Art and Research Needs.* Washington, DC: The National Academies Press.

Newman, Marissa. 2023. "Israel's new military AI systems select targets and plan missions 'in minutes.'" *The Japan Times,* July 17. www.japantimes.co.jp/news/2023/07/17/world/israel-quietly-embeds-ai-military-systems/.

Petch, Jeremy, Shuang Di and Walter Nelson. 2022. "Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology." *Canadian Journal of Cardiology* 38 (2): 204–13. https://doi.org/10.1016/j.cjca.2021.09.004.

Puscas, Ioana. 2022. "Confidence-Building Measures for Artificial Intelligence: A Framing Paper." Geneva, Switzerland: United Nations Institute for Disarmament Research. https://unidir.org/wp-content/uploads/2023/05/Confidence-Building_Final.pdf.

Say, Mark. 2024. "Ada Lovelace Institute urges evaluations of advanced AI." UKAuthority. July 29. www.ukauthority.com/articles/ada-lovelace-institute-urges-evaluations-of-advanced-ai/.

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

———. 2023b. "America Can Win the AI Race." *Foreign Affairs,* April 4. www.foreignaffairs.com/united-states/ai-america-can-win-race.

Schraagen, Jan Maarten. 2023. "Responsible use of AI in military systems: prospects and challenges." *Ergonomics* 66 (11): 1719–29. https://doi.org/10.1080/00140139.2023.2278394.

SIPRI. 2015. "LAWS at the CCW: Transparency and information sharing measures." April 17. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2015)/20150416_CCW_TRANSPARENCY.pdf.

Taddeo, Mariarosario, David McNeish, Alexander Blanchard and Elizabeth Edgar. 2019. "Ethical Principles for Artificial Intelligence in National Defence." *Philosophy & Technology* 34: 1707–29. https://doi.org/10.1007/s13347-021-00482-3.

Varella, Laura and Ray Acheson. 2024. "General Statements." *CCW Report* 12 (1): 3–8. https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2024/gge/reports/CCWR12.1.pdf.

Vergun, David. 2023. "U.S. Endorses Responsible AI Measures for Global Militaries." US Department of Defense. November 22. www.defense.gov/News/News-Stories/Article/Article/3597093/us-endorses-responsible-ai-measures-for-global-militaries/.

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

Yordanova, Tsvetelina. 2015. "The Transparency–Security Dilemma in National and International Context (A Comparative Analysis of the UN and NATO's Transparency/Secrecy Policies)." A paper presented at the Fourth Global Conference on Transparency Research in Lugano, Switzerland, June 4–6.

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