GC REAIM
GLOBAL COMMISSION ON RESPONSIBLE
AI IN THE MILITARY DOMAIN

TRENDS
تريندز للبحوث والاستشارات
TRENDS RESEARCH & ADVISORY

Foreign, Commonwealth
& Development Office

GC REAIM Expert Policy Note Series

An Approach for Assessing Autonomous and AI-Enabled Capabilities within Weapons Systems

Ariel Conn, Zena Assaad, and Catherine Tessier

May 2025

POWERED BY The Hague Centre for Strategic Studies

GC REAIM Expert Policy Note Series

An Approach for Assessing Autonomous and AI-Enabled Capabilities within Weapons Systems

Authors: Ariel Conn, Zena Assaad, and Catherine Tessier

May 2025

Cover photo: unsplash

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

The GC REAIM Expert Policy Note Series was funded by the Foreign, Commonwealth and Development Office (FCDO) of the United Kingdom. GC REAIM Experts maintained full discretion over the topics covered by the Policy Notes. The contents of the GC REAIM Expert Policy Note series do not represent the views of the Global Commission as a whole. The Policy Notes are intended to highlight some key issues around the governance of AI in the military domain and provide policy recommendations. This Policy Note was released during the Commission Meeting in Abu Dhabi, hosted and graciously supported by Trends Research and Advisory.

HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
@hcssnl

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

1. Introduction

Since the development of the United Nations (UN), certain classes of weapons have been more heavily regulated or outright banned, as member states of the UN deemed these weapons to be in violation of International Humanitarian Law. Examples of such weapons include chemical weapons, biological weapons, nuclear weapons, landmines, cluster munitions, and blinding lasers. With the rise of learning-based techniques–now most commonly referred to as “artificial intelligence (AI)”–along with other advanced computing capabilities, many have worried that weapons will increasingly function without sufficient human control or oversight. They argue that weapons that can function autonomously–that is, with minimal to no human control–represent a new class of weapons systems, which require new regulations, including bans, to ensure that responsible humans remain in control and accountable.

However, though the use of AI-based techniques and other advanced computing capabilities does represent a shift in warfare, autonomous functionality of weapons systems does not, in itself, reflect a new class of weapons. Weapons are typically classified in one of three ways: by the types of munition (e.g. conventional explosives versus nuclear, biological or chemical); by the delivery system (e.g. a bullet shot from a gun or a bomb dropped from a plane); or by the purpose of the system (e.g. anti-aircraft or heat-seeking missiles). Autonomous weapons systems (AWS), however, are weapons that have had autonomous capabilities designed into the system, regardless of the type of munition, delivery system, or the purpose of the weapon system.

Autonomous capabilities are ubiquitous in modern weapons systems. Weapons capabilities that can function autonomously include everything from navigation, to obstacle detection and avoidance, to identifying targets, to assisted decision-making, and much more. Weapons systems can include many different autonomous capabilities, and the impact of the capability will be affected by many different factors. These systems are enabled by an amalgam of internal components, including code, sensors and algorithmic functions that exist within the system itself, and external components, which cover a range of external dependencies, such as external networks and inputs or prompts from users. Because the interactions of these components are linked, an error, malfunction or unexpected outcome in one may have a domino effect on another. This complexity increases with the number of components a system has. Determining the cause and effect of an error, malfunction or unexpected output is also difficult because the interlaced dependencies of a system are not always related in direct or linear ways. The complexity of these systems means that it is difficult to determine and define when different autonomous components shift a weapons system from being categorised as “human controlled” to “autonomous.”

Moreover, whether or not a system is autonomous does not define or describe whether it is problematic. Some autonomous functions, such as path planning for a robot, are generally considered noncontroversial, while others are deeply controversial, such as the use of facial recognition to identify potential human targets. In fact, some uses of autonomy and/or AI in non-weaponised military systems may still be deeply controversial, create new risks or possibly even lead to unintended deaths. Alternatively, the use of AI in decision support systems may not qualify as “autonomous weapons” because humans are still in charge and overseeing the system, yet the use of the system may lead to a significant increase in civilian deaths. Focusing on “lethal autonomous weapons systems” may result in risky systems causing harm because they are overlooked in legal and regulatory efforts.

Additionally, a distinction must be made between automating *functions* versus the *techniques* that are used to do so. Specifically, concerns are often raised about the use of AI in AWS, however, it is not because a function is programmed with AI techniques that it is ethically problematic. For example, path planning for the example robot above might include a learning function to enable obstacle recognition. Conversely, automated functions may be ethically problematic without being programmed with AI techniques, such as anti-personnel mines. In fact, one of the challenges that has plagued the international debates around AWS is that most definitions of autonomous weapons can apply to some existing weapons systems that have already been deemed legal.

2. Reframing the Use of AI Techniques and Autonomy in Military Systems

Because autonomous capabilities can be added to or designed into virtually any weapons system, AWS cannot be meaningfully considered–or regulated–as a class of weapons systems. Considerations about the weapons systems need to focus on the *capabilities* rather than focusing only on the whole weapons systems. Questions to be asked should include: Which functions could be considered autonomous in the device? Which kinds of techniques were used to develop those functions? How do these autonomous capabilities interact to increase the complexity and variety of outcomes for the system? Who was involved in the development of–and thus may be responsible and accountable for–the different autonomous capabilities across the full lifecycle of the system?

Another issue arises when international discussions focus on AWS as a class of weapons systems: the vagueness of the framing exacerbates confusion and hyperbole around the abilities of various weapons systems. AWS are often defined as weapons that can "select and apply force to targets without human intervention."[1] They are also often referred to as "AI-enabled" or "artificially intelligent systems" or "autonomous and intelligent systems" or even "lethal autonomous weapons systems." These different terms are often used interchangeably, especially "AI weapons" and "autonomous weapons," though AI and autonomy are not synonymous. Moreover, the emphasis on AWS acting without human intervention can give the impression that these systems are "thinking machines" that decided to launch an attack on their own. In order to minimize confusion, definitional disagreements and AI hype, these discussions need to be reframed to focus more granularly on the specific autonomous capabilities that are more easily defined.

Rather than focusing on "autonomous weapons systems," "autonomy," or "artificial intelligence," this paper recommends reframing the discussions to focus on the specific and potentially problematic autonomous capabilities that can be built into various military systems. This paper proposes a method for identifying the autonomous capabilities in a system, determining which capabilities may have greater risks, monitoring how different autonomous capabilities in a single system may interact in unexpected ways, and tracking the human decision makers at each stage of the lifecycle.

Defining the autonomous components of systems, identifying if those components are higher risk and why, and addressing specific issues based on individual components is not a new approach. Lessons can be learned from related fields, such as aviation. The

[1] International Committee of the Red Cross. *ICRC Position on Autonomous Weapon Systems*. Geneva: ICRC, 12 May 2021. https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems.

aviation industry has long operated by focusing on the holistic lifecycle models of an aircraft. An aircraft is developed, certified, implemented, and maintained within a holistic system context, which, in addition to the technical components of the aircraft, encompasses the regulators, policy makers, engineers, manufacturers, pilots, passengers, etc. An aircraft is seldom examined independently of this broader system. If an incident occurs, such as a mid-air collision, an investigation is conducted across the entire lifecycle of that system, to find the technical point(s) of failure and the humans responsible. The primary points of failure may have occurred at earlier stages in the lifecycle even as secondary, catastrophic failures occurred at the time of use. For autonomy and AI in military systems, a similar approach should be adopted. These systems also exist within a holistic lifecycle that includes many different actors making critical decisions about various components of the system across the lifecycle, from research and development through to validation, implementation and use. A military system is a product of that lifecycle and cannot be regulated without knowledge and awareness of both the people and parts involved.

Two useful tools for conceptualizing this holistic lifecycle are the chain of responsibility (COR) model for human-machine teaming and the IEEE-SA's "A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defence Applications."[2] The COR imposes a primary duty on every person interacting with a system to reasonably ensure the safety of their activities. For any incident, an investigation is conducted that examines the entire lifecycle to determine where that duty was breached and by whom. The IEEE-SA Lifecycle Framework looks more granularly at the full lifecycle of the military system to define where major human decision-making nodes occur and to holistically address major legal, ethical, and technical challenges in the development and use of autonomy and AI in the system. These two tools are the basis for the recommendation made in this paper that a new database be developed to analyse the different autonomous capabilities.

[2] Brendan Walker-Munro and Zena Assaad, 'The Guilty (Silicon) Mind: Blameworthiness and Liability in Human-Machine Teaming', *Cambridge Law Review* 8, no. 1 (27 March 2023): 1–24; Sten Allik et al., 'A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications', *A Framework for Human Decision-Making through the Lifecycle of Autonomous andIntelligent Systems in Defense Applications*, October 2024, 1–63.

3. An Autonomous Capabilities Database

The primary database will be comprehensive, detailing autonomous capabilities that can be added to or designed into weapons and other defence systems and providing initial risk assessments for different autonomous capabilities. With developmental support by policy makers, this database would be publicly available, remaining a living document that will have new technologies and capabilities added as they are developed.

The following categories are recommended for the international database:

1. **The capability/tool:** Name and technical description of the capability. Variations of the capability may also be included here.
2. **Intended use case(s):** List of ways the tool is intended to and/or may be used.
3. **Types of programming:** List of the programming techniques that are used to design the capability (e.g., Kalman filter, constraint programming, supervised machine learning, LLMs, etc.).
4. **Known risks:** List risks and concerns about the capability, flagging issues considered high risk.

Meanwhile, developers of individual defence systems would create a more extensive database. This would highlight the specific autonomous capabilities in the system and outline human responsibility and accountability for these capabilities across the full lifecycle of the system. This database can ensure all members of the COR are familiar with how the autonomous capabilities function individually and together, and it can be reviewed throughout the lifecycle of the system as a means of ensuring humans remain aware of and responsible and accountable for the full capabilities and activities of each defence system. Most likely this database would remain relatively confidential or secret, though developers and/or countries are encouraged to contribute to updating the international database if the systems they develop exhibit new capabilities, or if known capabilities are applied to new use cases, built with different types of programming, or introduce new risks.

The database for the individual AWS should include the following categories:

1. **The capability/tool:** Name and technical description of the capability. Variations of the capability to be included here if they may be used.
2. **Intended use case(s):** List the ways the tool is expected to be used. If the intended use changes, this should be updated.
3. **Types of programming:** List types of AI techniques and other advanced computing technologies that are employed.
4. **Known and anticipated risks:** List which risks from the international database are applicable to this system. If other risks may arise as a result of this capability being paired with other capabilities in this particular system, list those here as well.

5. **Internal components:** An outline of the internal components of the system, e.g. code, sensors and algorithmic functions. There may be overlap here with #3, but many elements should also be new.
6. **External components:** An outline of the external components of the system, e.g. external networks and inputs or prompts from users.
7. **Design, research and development:** What risks might arise here? What should humans at this stage be watching for or aware of? What legal and ethical issues need to be considered? How is this capability expected to interact with the other capabilities of the full system?
8. **Procurement, acquisition and manufacturing:** What risks might arise here? What should humans at this stage be watching for or aware of?
9. **Testing, evaluation, verification and validation (TEVV):** What risks might arise here? What should humans at this stage be watching for or aware of? In addition to confirming that the capability acts as expected, how does it interact with other capabilities?
10. **Human training:** What do the human users, operators and commanders need to be trained on to use the capability within this system properly? What limitations do they need to be aware of? How might they misuse the system, and how can this be prevented?
11. **Tactical Deployment:** What risks might arise here? What should humans at this stage be watching for or aware of? If this system is designed within one country, can it legally be deployed in another?

As mentioned above, these categories are based on the COR model and the IEEE-SA Lifecycle Framework, however, the authors of this paper consider this a starting point, rather than the final list of categories. For example, the European Commission's paper, "How to complete your ethics self-assessment," also provides useful guidance.[3] Final development of the database may also be inspired by official documentation of the International Civil Aviation Organization, the European Union Aviation Safety Agency, and the US Federal Aviation Administration, along with existing risk management frameworks for militaries and weapons systems. Most important is the recommendation that the international community consider this approach to better define, understand, and address issues that arise as a result of using AI and autonomous capabilities in military systems.

[3] European Commission, *EU Grants: How to Complete Your Ethics Self-Assessment*, Version 2.0 (Brussels: European Commission, 13 July 2021), https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf.

4. Addendum: Facial Recognition Example

This addendum presents an example of how this model would work using facial recognition as the autonomous capability in question. It is an example and not intended to be a comprehensive review of all possible uses of facial recognition in military systems. Additionally, although the recommendation is to create a database, this example is presented as a list for the purposes of this paper.
Facial recognition example for the international database:

1. **The capability/tool:** Facial recognition. Similar tools include postural recognition and behavioural recognition.
2. **Intended use case(s):** Used to identify targets and their location. Used to survey potential threats. Used to identify friendly soldiers.
3. **Types of programming:** Machine learning, signals collected by sensors are processed by on-board or off-board software, reference databases can be built from field intelligence, open source intelligence (OSINT), digital social networks, etc.
4. **Known risks and concerns:** Automated targeting based on facial, postural or behavioural recognition could involve: targeting a particular person through authentication (e.g. a firing system that would trigger for this person); targeting a particular person in a public place through identification (e.g. an opportunity target identified by a loitering munition); or targeting all people with defined characteristics, through classification (e.g. all people wearing glasses). Inaccurate target recognition which, if unverified, can lead to engaging the wrong people.

Facial recognition example for an individual defence system:

1. **The capability/tool:** An automated targeting system based on facial, postural or behavioural recognition
2. **Intended use case(s):** Designed to target people based on their faces, body postures or characteristics, or dynamics of movements.
3. **Types of programming:** The signals collected by the sensors are processed by on-board or off-board software to interpret them for the targeting objective. For machine learning based recognition, reference databases can be built from field intelligence, open source intelligence (OSINT), digital social networks, etc.
4. **Known and anticipated risks:** Processing of human targets without human validation. Misuse of civil technologies for the general public (e.g., misuse of facial recognition to unlock smartphones or to access restricted areas). Poor performance and firing mistakes if the recognition system gives false negatives and false positives. A facial, postural or behavioural recognition system coupled with a sophisticated navigation system onboard a drone might allow the drone to more effectively target a terrorist, or it might allow a dictator to more effectively terrorize citizens.

5. **Internal components:** Hardware includes sensors such as visible or infrared cameras, radars, microphones, etc. The aim of the processing can be authentication, identification or categorization. The recognition process can be entirely automated or provide decision-aid for professionals.
6. **External components:** The external components of the system include cloud storage for data, computing infrastructure. Signals from sensors can be networked.
7. **Design, research and development:** The datasets used for developing and training the AI model may not be thorough enough to accurately capture different people. Why should this system be designed and used? Is there a need for more accurate weapons to target individuals? Is there a need to target people instead of facilities? Does this allow for better discrimination of targets and thus reduce collateral damage? Does it reduce reaction time of operators/soldiers/snipers?
8. **Procurement, acquisition and manufacturing:** System requirements at the point of negotiation may differ to those at the point of design. Alterations to the system may be required, potentially opening up opportunities for system vulnerabilities.
9. **Testing, evaluation, verification and validation (TEVV):** If the dataset is not granular enough, the testing for the system will be limited. Greater granularity of datasets is more time consuming and costly, therefore it is a trade-off. Is the recognition system capable of distinguishing that the targeted person is out of action, wounded, or surrendering? Is the recognition system capable of assessing context and what surrounds the targeted person? Such a system may induce people to modify their appearance and behaviour (disguise, camouflage, decoy, etc.). Is the system robust to such modifications? How is uncertainty about recognition dealt with? On which databases is the recognition system based (e.g. military or intelligence databases, or general databases built from the digital social networks)? What does a recognition rate of 96% mean exactly? How are the databases and the system tested, validated and updated? How are biases dealt with in the databases and in the system's architecture, parameters, thresholds, or machine learning models?
10. **Human training:** To what extent does the result of the recognition system influence the human decision to fire? Is there room for doubt, for instance when the human supervisor is uncertain about the person's identity or behaviour whereas the recognition system gives a clear result? Conversely, the human supervisor may overly trust the result of the recognition system. Training is necessary to verify system outputs and identify incorrect outputs.
11. **Tactical Deployment:** Human decision must take place *before* the actual recognition on the field, which presupposes that the context does not change between the decision and the action. The system will need to be frequently updated, particularly in dynamic operating environments. If deployed domestically, the system will need to meet domestic regulations and will need to include a dataset that is reflective of that domestic environment. If deployed internationally, the system may require specific changes to meet international regulations. Datasets may also need to be adjusted to better reflect the operating environment.

About the Authors

Ms. Ariel Conn

Ariel Conn leads the IEEE-SA Research Group on Issues of Autonomy and AI for Defence Systems. She's working with a group of experts to develop a more nuanced framework to address ethical and technical challenges of AI and autonomy in defence systems, as well as ensure human responsibility and accountability across every stage of a defence system's lifecycle. Publications from this effort include the Ethical and technical challenges in the development, use, and governance of autonomous weapons systems.

Dr. Zena Assaad

Dr Zena Assaad is a senior lecturer in the School of Engineering at the Australian National University and is also a fellow with the Australian Army Research Centre. She has previously held a fellowship with Trusted Autonomous Systems. Her research explores the safety of human-machine teaming and the assurance and certification of trusted autonomous and AI systems. Dr Assaad is the founder and chair of the Australian national community of practice for UAS and AAM research.

Dr. Catherine Tessier

Dr. Catherine Tessier is a Director of Research at ONERA in Toulouse, France, and ONERA's Research Integrity and Ethics Officer. Her research focuses on the modelling of ethical frameworks, on ethical issues related to the “autonomy” of robots and on digital ethics. She is a member of the French National Committee for Digital Ethics and a member of the French Defence Ethics Committee. She was a member of the UNESCO ad hoc Expert Group for the elaboration of the Recommendation on the Ethics of Artificial Intelligence.

HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
@hcssnl

The Hague Centre for Strategic Studies