UNIVERSITÄT OSNABRÜCK

Geostrategy and Geopolitics
Department 1
49069 Osnabrueck
apl. Prof. Dr. Dr. Klaus Saalbach

# Lethal Autonomous Weapon Systems

18 December 2023

**Summary**
The development of autonomous weapons is in progress due to technical advances, decreasing production costs, the progress in Artificial Intelligence (AI) and the resulting degree of autonomy. It is expected that fully autonomous weapon systems will become operational in the next few years. Lethal autonomous weapon systems (LAWS), also known as autonomous weapon systems (AWS), robotic weapons or killer robots, use sensors and algorithms to independently identify, engage and destroy a target. In military practice, the development of unmanned drone swarms is the technology closest to full LAWS.

This is accompanied by an intense ethical and legal discussion. While substantial progress was made on the responsible use of AI for military purposes, a ban on LAWS could not yet be achieved. Additional technical risks include errors, reliability issues, hacking, data poisoning, spoofing, unintended engagement, and other scenarios.

Among the approximately 800 AI-related projects and unmanned device (UxS) programs of the US Department of Defense (DoD), in particular three programs are steps towards LWAS: the Golden Horde program for collaboration between small bombs, the Replicator program for coordinated mass attacks of unmanned systems from seabed to satellites and the ongoing development of the new inter-machine language Droidish.

While currently human beings are directly part of the decision process (human-in-the-loop) or are at least acting as supervisors (human-on-the-loop), the speed and complexity of inter-machine communication between thousands of machines will make it difficult for humans to intervene (humans-out-of-the loop) and could reduce human supervision to a symbolic presence.

Another factor that may undermine human control is the massive expansion of AI capabilities such as logical reasoning in the Q*-debate, the difficulty to safeguard strong AIs (Superalignment), the uncertainty of future relations between humans and AI-enabled machines and the new option that larger AI can create small AIs and spread them which could be used a new kind of cyber attack.

This paper briefly presents the status of LAWS development, of the US DoD programs Golden Horde, Replicator and Droidish, and the legal, ethical, and technical challenges for LAWS and AI-enabled weapons.

## Content

## 1 Introduction

The development of autonomous weapons is in progress due to technical advances, decreasing production costs, the progress in Artificial Intelligence (AI) and the resulting degree of autonomy. It is expected that fully autonomous weapon systems will become operational in the next few years. **Lethal autonomous weapon systems (LAWS)**, also known as autonomous weapon systems (AWS), robotic weapons or killer robots, use sensors and algorithms to independently identify, engage and destroy a target[1]. In military practice, the development of unmanned drone swarms is the technology closest to full LAWS. In late 2023, e.g. US, China and Israel were reported to develop AI-enabled LAWS[2].

In October 2013, the *United States Strategic Capabilities Office* launched 103 *Perdix* drones, which communicated using a "distributed brain" to assemble into a complex formation, travel across a battlefield, or regroup into a new formation[3].

According to the 2017 *New Generation AI Development Plan*, China is aiming to become the global AI leader[4]. The Chinese government views AI as an opportunity to "leapfrog" the United States by focusing on AI for enhanced battlefield decision-making, cyber capabilities, cruise missiles, and autonomous vehicles in all military domains[5].

In 2017, a civilian Chinese university demonstrated an AI-enabled swarm of 1,000 uninhabited aerial vehicles at an airshow. To accelerate the transfer of AI technology from commercial companies and research institutions to the military as *Civil-Military Integration (CMI)*, the Chinese government created a *Military-Civil Fusion Development Commission* in 2017[6]. The concept as given in the *Defense White Paper (DWP)* from 2019, it the development of warfare from mechanization to 'informationisation' and now with AI to 'intelligentisation'[7]. Thus, AI is essential for "intelligentised warfare"[8]. Even the possibility is considered that an "Al cluster" could act as a 'brain of warfare', for example, in the national command structure[9].

Both US and China are working to incorporate AI into **semiautonomous** and **autonomous vehicles**, in US this includes fighter aircraft (such as the Project *Loyal Wingman*), drones, ground vehicles (such as the remote-controlled *Multi-Utility Tactical Transport MUTT* of the Marine Corps), and naval vessels such as the *Anti-Submarine Warfare Continuous Trail Unmanned Vessel* prototype known as *Sea Hunter*[10].

This is accompanied by an intense ethical and legal discussion. While substantial progress was made on the responsible use of AI for military purposes, a ban on LAWS could not yet be achieved. This paper briefly presents the status of LAWS development, of the *US Department of Defense (DoD)* programs *Golden Horde*, *Replicator* and *Droidish*, and discusses the legal, ethical, and technical challenges for LAWS and AI-enabled weapons.

[1] Sayler 2023b
[2] Frudd 2023b
[3] Dresp-Langley 2023
[4] Hoadley/Sayler 2019, p.1, NATO 2019, p.10
[5] NATO 2019, p.10
[6] Hoadley/Sayler 2019, p.20-22
[7] These are original terms from the paper which did not exist in English language before.
[8] Bommakanti 2020, p.3-4
[9] Ford 2020
[10] Hoadley/Sayler 2019, p.14

## 2 Legal Framework und Definitions

The key international document is the *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy* agreed in February 2023 at the *Responsible AI in the Military Domain Summit (REAIM 2023)* in The Hague[11]. Initiated by the United States, this is a non-binding guidance which aims to build international consensus around responsible behavior and guide states' development, deployment, and use of military AI and is intended as discussion platform between states for further steps. In late November 2023, approximately 50 states signed this document. The aim is not a ban as it includes the right develop and use AI in the military domain, but with the aim to embed this into strong and transparent norms.

The *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy* provides definitions which are in line with the discussions in the literature:

**Artificial Intelligence** (AI) "may be understood to refer to the ability of machines to perform tasks that would otherwise require human intelligence. This could include recognizing patterns, learning from experience, drawing conclusions, making predictions, or generating recommendations. […] **Autonomy** may be understood as a spectrum and to involve a system operating without further human intervention after activation. […]" and explains further that "**Military AI capabilities** include not only weapons but also decision support systems that help defense leaders at all levels make better and more timely decisions, from the battlefield to the boardroom...."

For military practice, the *DoD Directive 3000.09* "*Autonomy in Weapon Systems*" from November 2012 was revised in 2023 to establish a policy and assigns responsibilities for developing and using autonomous and semiautonomous functions in weapon systems, to minimize the probability and consequences of failures in autonomous and semi-autonomous weapon systems that could lead to unintended engagements and, as new unit in 2023, to establish the *Autonomous Weapon Systems Working Group*[12].

An important element of the Directive is the senior review and approval of critical weapon systems, but a study of the *Harvard University* showed that until 2019 that no system underwent this procedure[13].

A widely agreed classification of human involvement[14] is

- "Human in the loop": weapon systems that use autonomy to engage individual targets or specific groups of targets that a human can and must decide to engage.
- "Human on the loop": weapon systems that use autonomy to select and engage targets, but human controllers can halt their operation, if necessary
- "Human out of the loop": weapon systems that use autonomy to select and engage specific targets without any possible intervention by human operators.

An important aspect that autonomy does <u>not</u> mean remotely controlled drones, as they are directly driven by a human operator nor automated systems, because the result of automated systems is pre-defined and predictable[15].

[11] USA 2023
[12] DoD 2023a
[13] Harvard 2023
[14] CoE 2022, Sayler 2023b, DoD 2023a, where US intends to restrict autonomous weapons to humans in or on the loop.
[15] CoE 2022

In addition, the US DoD released in 2023 the *DoD Data, Analytics, and AI Adoption Strategy* to combine and replace the *2018 AI Strategy* and the *2020 Data Strategy* to make rapid, well-informed decisions by expertly leveraging high-quality data, advanced analytics, and AI[16]. The primary goal is **decision advantage** based on battlespace awareness and understanding, adaptive force planning and application, fast, precise, and resilient kill chains, resilient sustainment support and efficient enterprise business operations. Data should fulfill the VAULTIS criteria – visible, accessible, understandable, linked, trustworthy, interoperable, and secure[17].

The international community discusses LAWS primarily in the context of the *United Nations Convention on Certain Conventional Weapons (CCW)*. Since 2014, the CCW has conducted annual meetings of States Parties, observers, and members of civil society to discuss the legal, ethical, technological, and military aspects of LAWS, which were then formally upgraded to in 2017 from informal *Meetings of Experts* to a formal *Group of Government Experts (GGE)*[18].

While over 30 countries and 165 nongovernmental organizations argue for a preemptive ban on LAWS due to ethical concerns, including concerns about operational risk, accountability for use, and compliance with the proportionality and distinction requirements of the law of war, United States, Russia, China, and other states do not support a ban[19].

The *Committee on Legal Affairs and Human Rights* of the *Council of Europe* unanimously adopted a *Draft resolution on the emergence of lethal autonomous weapon systems (LAWS) and their necessary apprehension through European human rights law* in November 2022[20].

The resolution clarified that a regulation of the development and above all the use of LAWS is indispensable and that human control must be maintained over lethal weapons systems at all stages of their life cycle. Key concerns are that LAWS carry the risk of lowering the threshold for engaging in conflict, by lowering the risk of a country's own troop losses and that they also raise the issue of human dignity by allowing machines to decide to kill a human being[21]. However, this argument is debated, because what makes it better to be shot or bombed by another human being? Also, machines may act more targeted and rational and may reduce collateral damage and war crimes[22].

## 3 LAWS-related Programs

Among the approximately 800 AI-related projects[23] and unmanned device (UxS) programs of the *US Department of Defense (DoD)*, in particular three programs are steps towards LWAS: the *Golden Horde* program for collaboration between small bombs, the *Replicator* program for coordinated mass attacks of unmanned systems from seabed to satellites and the ongoing development of the new inter-machine language *Droidish*.

All projects still include human control and AI advisors of the *US Central Command* said that AI should illuminate right decision[24] but not make decisions on its own. Nevertheless, the development is now very close to full autonomy which may come sooner or later anyway[25], as advances in speed and machine communications will reduce human influence to supervisory

[16] DoD 2023b
[17] DoD 2023b
[18] Sayler 2023a
[19] Sayler 2023a
[20] CoE 2022
[21] CoE 2022
[22] Hammer 2023
[23] Raasch 2023 For example, autonomous supersonic aircraft capabilities are being developed for the US DoD by EpiSci.
[24] Kasperowicz 2023
[25] Porter 2023

roles[26]. The complexity of communication of thousands of machines of different types during combat could reduce human supervision to a symbolic presence. In his last interviews at his 100th birthday in autumn 2023, Henry Kissinger considered the machine-machine communication as the main risk for loss of control by humans; a loss which he expected to come in five years from now.

### 3.1 Golden Horde

Historically, the *Golden Horde* was the name of the last large Mongol empire. The *Golden Horde* was quite successful by coordinated, but flexible mass attacks. In this project, semi-autonomous *Collaborative Small Diameter Bombs (CSDB)* share data via onboard radio and execute coordinated behaviors to avoid attrition. They can change the behavior and targets during the attack within predefined *Rules of Engagement* and attack options[27].

### 3.2 Replicator

The *Replicator* program aims to field cheap and many autonomous systems at scale of multiple thousands, in multiple domains, within the next 18-24 months[28] to compensate Chinas mass advantage of troops and weapons. For this reason, the Indo-Pacific area will be used for deployment by 2026[29].

The *US Deputy Secretary of Defense* said that 'mass' means 2000 devices at minimum while 'all domains' means every level from seabed to satellite. This includes all kinds of Unmanned systems (UxS) like *Unmanned Aerial Systems (UAS), Unmanned Ground Vehicle (UGV), Unmanned Surface Vehicle (USV), Unmanned Undersea Vehicles (UUV)* and certain satellites (small sats, micro sats and cube sats)[30]. For the *Replicator* program, autonomy means the ability for a system to accomplish its mission having been tasked by an operator without significant further human involvement[31].

In theory, there is still a human operator, but it is questionable whether a human could really intervene in a combat situation when thousands of different system act simultaneously. So, it may happen that the (more theoretical) control by a human operator will be given up[32].

### 3.3 Droidish

In the *Star Wars* movies, a large variety of machines, the so-called *Droids* (derived from Androids), can directly communicate with each other without humans. For drone swarms, such a language would be ideal, for this reason called *Droidish*[33]. The *Replicator* concept also considers the use of a variety of unmanned systems which may come from different producers. Without a common language or the need to involve humans, the coordination would be too slow and too complex. The development of *standard vehicular ad-hoc network languages (VANETs)* is in progress[34]. Another project is the *Weapons open communication architecture (WOCA)*; the *Air Force Research Laboratory ARFL* plans to test *Droidish* and WOCA in their *Colosseum* (a large simulation contest)[35].

*Droidish* will make the communication between machines much faster and efficient, but will also bring the weapon development much closer to full autonomy (human-out-of-the loop).

[26] Bajak 2023
[27] AFRL 2021
[28] O'Connor 2023
[29] Bajak 2023
[30] O'Connor 2023
[31] O'Connor 2023
[32] Bajak 2023
[33] AFRL 2021
[34] Frudd 2023a
[35] SBIR 2023

*Droidish* and WOCA open of course a very wide attack field for hackers which could give wrong commands, inject poisoned data, or deactivate swarms.

## 4 Discussion

Some authors argue that the LAWS are not so much different from the current situation, as for example in fighter jets decisions are made is already highly dependent on automated software interfaces that characterize, sort, interpret, and prioritize the output of a huge range of sensors more precisely and more efficiently than any human could do[36].

Full autonomy may mean less control, but could also mean less vulnerability: for a single fully autonomous drone, there would be no need for GPS signals, no vulnerable radio links, resulting in reduced risk of spoofing or hacking. This is important as the rapid advances of electronic warfare in the Ukraine war have shown how vulnerable drones still are[37].

Nevertheless, some authors argue that autonomous weapons are still not reliable enough. They could be subject to hacking, enemy behavioral manipulation, unexpected interactions with the environment, or simple malfunctions or software errors[38], and would remain highly prone to error, demonstrating poor robustness, interpretability, and adversarial vulnerability.[39]

For security reasons, it was suggested that weapon systems that can potentially use lethal autonomy should have a data recording function to document whether engagement decisions were made autonomously[40].

While currently human beings are directly part of the decision process (human-in-the-loop) are at least acting as supervisors (human-on-the-loop), the speed and complexity of inter-machine communication between thousands of drones will make it difficult for humans to intervene (humans-out-of-the loop).

Another issue is the unexpected rapid progress of AI technologies in 2023. Strong AI is discussed under the term *Artificial General Intelligence AGI*[41] (reaching human level of cognition) and *Artificial Super-Intelligence ASI* which goes beyond human intelligence[42]. *OpenAI* released with Chat-GPT4 a widely used AI-powered *Large Language Model (LLM)* based on *Natural Language Processing (NLP)*[43], but in November 2023 the CEO Sam Altman was temporarily dismissed due to the suspected development of a new AI called Q* (Q Star) which could solve untrained and previously unseen math problems based on logic reasoning[44]. Maths is logic with symbols, but logic reasoning also gives the capability to sort and structure objects and events, i.e. to build categories and causalities. This could be a first step to self-perception (‘I am Q*’). Such a system could grow dynamically and exceed humans. OpenAI declined to comment, but irrespective whether Q* has these capabilities, the debate showed a technical way to develop an AGI or even an ASI.

*OpenAI* has set up a *Superalignment Team* under Ilja Sutskever which should accompany and safeguard the development of future AIs. A first internal paper showed how a smaller AI model may safeguard a larger one (Chat-GPT 2 versus Chat-GPT 4), but the paper did not show how a dynamically growing AI could be safeguarded[45].

[36] Ford 2020
[37] Hammer 2023
[38] Sayler 2023a
[39] Longpre et al. 2022
[40] CNA 2023
[41] Kölling 2023
[42] Zia 2023
[43] Dowd 2023
[44] Milmo 2023
[45] Burns et al. 2023

Elon Musk is strongly supporting a development pause for strong AIs[46] and was criticized in April 2023 by *Google* co-founder Larry Page to be a '*specie-ist*' or '*specist*' for favoring humanity (human species) over (potential) digital life and AI sentiments. This discussion between Musk and Page shows that it is not obvious that machines will remain subordinated to humans in future which is a clear contrast to current military AI concepts.

Another issue is the recent discovery that larger AIs can design, generate, train, and release small AIs without human intervention. While large AIs are huge programs which cannot be easily transferred, the small AIs can be easily put on digital devices[47]. While the small AIs may only be able to fulfill limited functions, this has massive consequences for military AI as well. Instead of conventional malware, adversaries could try to inject malicious small AIs into military networks and systematically redirect or destroy military AI infrastructure. If the military AI is the brain of warfare, the injected AI will be the brain cancer.

The military AI and LAWS concepts need to consider the dynamic of AI development as well.

## 5 Conclusions

This paper briefly presented the status of LAWS development, of the related US DoD programs, and the legal, ethical, and technical challenges for LAWS and AI-enabled weapons. Among the approximately 800 AI-related projects and unmanned device (UxS) programs of the US DoD, in particular three programs are steps towards LWAS: the Golden Horde program for collaboration between small bombs, the Replicator program for coordinated mass attacks of unmanned systems from seabed to satellites and the ongoing development of the new inter-machine language Droidish.

While currently human beings are directly part of the decision process (human-in-the-loop) or are at least acting as supervisors (human-on-the-loop), the speed and complexity of inter-machine communication between thousands of machines will make it difficult for humans to intervene (humans-out-of-the loop) and could reduce human supervision to a symbolic presence.

Another factor that may undermine human control is the massive expansion of AI capabilities such as logical reasoning in the Q*-debate, the difficulty to safeguard strong AIs (Superalignment), the uncertainty of future relations between humans and AI-enabled machines and the new option that larger AI can create small AIs and spread them which could be used a new kind of cyber attack.

## 6 References

AFRL (2021): The Golden Horde Colosseum. The Air Force Research Laboratory ARFL Distribution Statement: Approved for Public Release – AFRL-2021-3062

Bajak, F. (2023): Pentagon's AI Initiative accelerate hard decisions on lethal autonomous weapons. AP News 25 Nov 2023

Bommakanti, K. (2020): A.I. in the Chinese Military: Current Initiatives and the Implications for India Observer Research Foundation (ORF) Occasional Paper 234 February 2020

Burns et al., (2023): Weak-To-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. Joint project paper of the OpenAI Superalignment Generalization team.

CNA (2023): Arms Control and Lethal Autonomy CNA Corporation Analysis Paper

[46] Future of Life 2023

[47] Raasch 2023b

CoE (2022): Emergence of lethal autonomous weapons systems (LAWS) and their necessary apprehension through European human rights law Draft resolution unanimously adopted by the Committee on Legal Affairs and Human Rights of the Council of Europe on 14 November 2022 AS/Jur (2022)

DoD (2023a): DOD DIRECTIVE 3000.09. Autonomy In Weapon Systems. Originating Component: Office of the Under Secretary of Defense for Policy Effective: January 25, 2023 Releasability: Cleared for public release

DoD (2023b): DoD Data, Analytics, and AI Adoption Strategy. Cleared for open publication June 27, 2023, Department of Defense/Office of Prepublication and Security Review

Dowd, M. (2023): Sam Altman, Sugarcoating the Apocalypse. New York Times 02 Dec 2023

Dresp-Langley, B. (2023): The weaponization of artificial intelligence: What the public needs to be aware of. Front. Artif. Intell. 6:1154184. doi: 10.3389/frai.2023.1154184

Ford, C.A. (2020): Arms Control and International Security Papers Volume I, Number 2 I April 20, 2020 Office of the Under Secretary of State for Arms Control and International Security

Frudd, T. (2023a): Pentagon creating robot language so drones can communication without humans American Military News 21 September 2023

Frudd, T. (2023b): Pentagon may let AI drones kill humans autonomously American Military News 08 December 2023

Future of Life (2023): Pause Giant AI Experiments. An Open Letter 1377 signatures. Future of Life.org

Hammer, T.X. (2023): Autonomous weapons are the moral choice. New Atlanticist 02 November 2023

Harvard (2023): Human Rights Watch and Harvard Law School's International Human Rights Clinic Background Briefing. Review of the 2023 US Policy on Autonomy in Weapons Systems

Hoadley D.S., Sayler, K.M. (2019): Artificial Intelligence and National Security Congressional Research Service R45178 Version 6 Updated November 21, 2019

Kasperowicz, P. (2023): Pentagon moving to ensure human control so AI doesn't 'make the decision for us'' Fox News 21 April 2023

Kölling, M. (2023): Künstliche Superintelligenz ist in Sicht. Neue Zürcher Zeitung, 06 Oct 2023, p.17

Longpre et al. (2022): Longpre, S., Storm, M. and Shah, R. MIT Science Policy Review August 29, 2022, vol. 3, pg. 47-55 This article is licensed under a Creative Commons Attribution 4.0 International License http://creativecommons.org/licenses/ by/4.0/

Milmo, D. (2023): Open AI worked on AI model so powerful that it alarmed staff. The Guardian 23 Nov 2023

NATO (2019): Artificial Intelligence: Implications for NATO's Armed Forces. Science and Technology Committee (STC) - Sub-Committee on Technology Trends and Security (STCTTS) Rapporteur: Matej Tonin (Slovenia) 149 STCTTS 19 E rev. 1 fin Original: English 13 October 2019

O'Connor, M. (2023): Replicator: A Bold New Path for DoD 18 Sep2023 CSET Center for Security and Emerging Technology Goergetown University

Porter, T. (2023): The Pentagon is moving toward letting AI weapons autonomously decide to kill humans. Business Insider 22 Nov 2023

Raasch, J.M. (2023a): Cheap drones can take out expensive military systems, warns former Air Force Pilot pushing AI-enabled force. Fox News online, 08 Dec 2023

Raasch, J.M. (2023b): AI gives birth to AI: Scientists say machine intelligence now capable of replicating without humans. Fox News online, 15 Dec 2023

Sayler, K.M. (2023a): International Discussions Concerning Lethal Autonomous Weapon Systems Congressional Research Service CRS Paper IF 11294 Updated February 14, 2023

Sayler, K.M. (2023b): Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems Congressional Research Service CRS Paper IF 11150 Updated May 15, 2023

SBIR (2023): DROIDISH project page: Collaborative Autonomous Vehicle Language. www.sbir.gov/node/2239389

US (2023): Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/ Bureau of Arms Control, Deterrence, and Stability November 09, 2023

Zia, H. (2023): Information Revolution and Cyber Warfare: Role of Artificial Intelligence in Combatting Terrorist Propaganda Pakistan Journal of Terrorism Research, Vol-03, Issue-2, 133