ARIS MESSINIS/AFP/GETTY

In the Donbas region of Ukraine, a Ukrainian soldier prepares a drone to carry a hand grenade for an attack in March 2023.

LETHAL AI WEAPONS ARE HERE: HOW CAN WE CONTROL THEM?

Weapons guided by artificial intelligence are already in use. Researchers, legal experts and ethicists are struggling with what should be allowed on the battlefield. **By David Adam**

In the conflict between Russia and Ukraine, video footage has shown drones penetrating deep into Russian territory, more than 1,000 kilometres from the border, and destroying oil and gas infrastructure. It's likely, experts say, that artificial intelligence (AI) is helping to direct the drones to their targets. For such weapons, no person needs to hold the trigger or make the final decision to detonate.

The development of lethal autonomous weapons (LAWs), including AI-equipped drones, is on the rise. The US Department of Defense, for example, has earmarked US$1 billion so far for its Replicator programme, which aims to build a fleet of small, weaponized autonomous vehicles. Experimental submarines, tanks and ships have been made that use AI to pilot themselves and shoot. Commercially available drones can use AI image recognition to zero in on targets and blow them up. LAWs do not need AI to operate, but the technology adds speed, specificity and the ability to evade defences. Some observers fear a future in which swarms of cheap AI drones could be dispatched by any faction to take out a specific person, using facial recognition.

Warfare is a relatively simple application for AI. "The technical capability for a system to find a human being and kill them is much easier than to develop a self-driving car. It's a graduate-student project," says Stuart Russell, a computer scientist at the University of California, Berkeley, and a prominent campaigner against AI weapons. He helped to produce a viral 2017 video called *Slaughterbots* that highlighted the possible risks.

The emergence of AI on the battlefield has spurred debate among researchers, legal experts and ethicists. Some argue that AI-assisted weapons could be more accurate than human-guided ones, potentially reducing both collateral damage — such as civilian casualties and damage to residential areas — and the numbers of soldiers killed and maimed, while helping vulnerable nations and groups to defend themselves. Others emphasize that autonomous weapons could make catastrophic mistakes. And many observers have overarching ethical concerns about passing targeting decisions to an algorithm.

For years, researchers have been campaigning to control this new threat[1]. Now the United Nations has taken a crucial step. A resolution last December added the topic of LAWs to the agenda of the UN General Assembly meeting this September. And UN secretary-general António Guterres stated last July that he wants a ban on weapons that operate without human oversight to be in place by 2026. Bonnie Docherty, a human-rights lawyer at Harvard Law School in Cambridge, Massachusetts, says that getting this topic on to the UN agenda is significant after a decade or so of little progress. "Diplomacy moves slowly, but it's an important step," she says.

The move, experts say, offers the first realistic route for states to act on AI weapons. But this is easier said than done. These weapons raise difficult questions about human agency, accountability and the extent to which officials should be able to outsource life-and-death decisions to machines.

Under control?

Efforts to control and regulate the use of weapons date back hundreds of years. Medieval knights, for example, agreed not to target each other's horses with their lances. In 1675, the warring states of France and the Holy Roman Empire agreed to ban the use of poison bullets.

Today, the main international restrictions on weaponry are through the UN Convention on Certain Conventional Weapons (CCW), a 1983 treaty that has been used, for example, to ban blinding laser weapons.

Autonomous weapons of one kind or another have been around for decades at least, including heat-seeking missiles and even (depending on how autonomy is defined) pressure-triggered landmines dating back to the US Civil War. Now, however, the development and use of AI algorithms is expanding their capabilities.

The CCW has been formally investigating AI-boosted weapons since 2013, but because it requires international consensus to pass regulations – and because many countries actively developing the technology oppose any ban – progress has been slow. In March, the United States hosted an inaugural plenary meeting on the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, a parallel effort that emphasizes voluntary guidelines for best practice rather than a legally enforceable ban.

Part of the problem has been a lack of consensus about what LAWs actually are. A 2022 analysis found at least a dozen definitions of autonomous weapons systems proposed by countries and organizations such as the North Atlantic Treaty Organization (NATO)[2]. The definitions span a wide range and show a limited amount of agreement on, or even an understanding of, AI, says Russell.

The United Kingdom, for example, says LAWs are "capable of understanding higher-level intent and direction", whereas China says such a weapon can "learn autonomously, expand its functions and capabilities in a way exceeding human expectations". Israel declares: "We should stay away from imaginary visions where machines develop, create or activate themselves – these should be left for science-fiction movies." Germany includes "self-awareness" as a necessary attribute of autonomous weapons – a quality that most researchers say is far away from what's possible with AI today, if not altogether impossible.

"That sort of means that the weapon has to wake up in the morning and decide to go and attack Russia by itself," says Russell.

Although a more comprehensive, specific and realistic definition for LAWs will need to be ironed out, some experts say this can wait. "Traditionally in disarmament law, although it's counter-intuitive, actually they often do the definition last in negotiation," Docherty says. A working definition is usually enough to start the process and can help to soften initial objections from countries opposed to action.

The AI advantage

According to a 2023 analysis published by the Center for War Studies at University of Southern Denmark in Odense[3], the autonomous weapons guided by AI available to army commanders today are relatively crude – slow-moving and clumsy drones equipped with enough explosive to blow up themselves and their targets.

> "THESE THINGS ARE GOING TO GET SHOT DOWN. THEY'RE GOING TO BE CAPTURED."

These 'loitering munitions' can be the size of a model aircraft, cost about $50,000, and carry a few kilograms of explosive up to 50 kilometres away, enough to destroy a vehicle or to kill individual soldiers. These munitions use on-board sensors that monitor optical, infrared or radio frequencies to detect potential targets. The AI compares these sensor inputs with pre-designated profiles of tanks, armoured vehicles and radar systems – as well as human beings.

Observers say that the most significant advantage offered by these autonomous bombs over remote-controlled drones is that they still work if the other side has equipment to jam electronic communications. And autonomous operation eliminates the risk that remote operators could be traced by an enemy and themselves attacked.

Although there were rumours that autonomous munitions killed fighters in Libya in 2020, reports from the conflict in Ukraine have cemented the idea that AI drones are now being used. "I think it's pretty well accepted now that in Ukraine, they have already moved to fully autonomous weapons because the electronic jamming is so effective," says Russell. Military commanders such as Ukraine's Yaroslav Honchar have said that the country "already conducts fully robotic operations, without human intervention"[3].

It's hard to know how well AI weapons perform on the battlefield, in large part because militaries don't release such data. Asked directly about AI weapons systems at a UK parliamentary enquiry last September, Tom Copinger-Symes, the deputy commander of the UK Strategic Command, didn't give much away, saying only that the country's military is doing benchmarking studies to compare autonomous with non-autonomous systems. "Inevitably, you want to check that this is delivering a bang for a buck compared with the old-fashioned system of having ten imagery analysts looking at the same thing," he said.

Although real-world battlefield data are sparse, researchers note that AI has superior processing and decision-making skills that, in theory, offer a significant advantage. In annual tests of rapid image recognition, for example, algorithms have outperformed expert human performance for almost a decade. A study last year, for example, showed that AI could find duplicated images in scientific papers faster and more comprehensively than a human expert[4].

In 2020, an AI model beat an experienced F-16 fighter-aircraft pilot in a series of simulated dogfights thanks to "aggressive and precise manoeuvres the human pilot couldn't outmatch". Then, in 2022, Chinese military researchers said that an AI-powered drone had outwitted an aircraft flown remotely by a human operator on the ground. The AI aircraft got onto the tail of its rival and into a position where it could have shot it down.

A drone AI can make "very complex decisions around how it carries out particular manoeuvres, how close it flies to the adversary and the angle of attack", says Zak Kallenborn, a security analyst at the Center for Strategic and International Studies in Washington DC.

Still, says Kallenborn, it's not clear what significant strategic advantage AI weapons offer, especially if both sides have access to them. "A huge part of the issue is not the technology itself, it's how militaries use that technology," he says.

AI could also in theory be used in other aspects of warfare, including compiling lists of potential targets; media reports have raised concerns that Israel, for example, used AI to create a database of tens of thousands of names

US AIR FORCE PHOTO/KYLE BRASIER

US Air Force aircraft have been used to test the capabilities of autonomous agents.

of suspected militants, although the Israeli Defence Forces said in a statement that it does not use an AI system that "identifies terrorist operatives".

Line in the sand

One key criterion often used to assess the ethics of autonomous weapons is how reliable they are and the extent to which things might go wrong. In 2007, for example, the UK military hastily redesigned its autonomous Brimstone missile for use in Afghanistan when it was feared it might mistake a bus of schoolchildren for a truckload of insurgents.

AI weapons can fairly easily lock on to infrared or powerful radar signals, says Kallenborn, comparing them to a library of data to help decide what is what. "That works fairly well because a little kid walking down the street is not going to have a high-powered radar in his backpack," says Kallenborn. That means that when an AI weapon detects the source of an incoming radar signal on the battlefield, it can shoot with little risk of harming civilians.

But visual image recognition is more problematic, he says. "Where it's basically just a sensor like a camera, I think you're much, much more prone to error," says Kallenborn. Although AI is good at identifying images, it's not foolproof. Research has shown that tiny alterations to pictures can change the way they are classified by neural networks, he says – such as causing them to confuse an aircraft with a dog[5].

Another possible dividing line for ethicists is how a weapon would be used: to attack or defend, for example. Sophisticated autonomous radar-guided systems are already used to defend ships at sea from rapid incoming targets. Lucy Suchman, a sociologist at Lancaster University, UK, who studies the interactions between people and machines, says that ethicists are more comfortable with this type of autonomous weapon because it targets ordnance rather than people, and because the signals are hard to falsely attribute to anything else.

One commonly proposed principle among researchers and the military alike is that there should be a 'human in the loop' of autonomous weapons. But where and how people should or must be involved is still up for debate. Many, including Suchman, typically interpret the idea to mean that human agents must visually verify targets before authorizing strikes and must be able to call off a strike if battlefield conditions change (such as if civilians enter the combat zone). But it could also mean that humans simply program in the description of the target before letting the weapon loose – a function known as fire-and-forget.

Some systems allow users to toggle between fully autonomous and human-assisted modes depending on the circumstances. This, say Suchman and others, isn't good enough. "Requiring a human to disable an autonomous function does not constitute meaningful control," she says.

The idea of full autonomy also muddies the water about accountability. "We're very concerned about the use of autonomous weapons systems falling in an accountability gap because, obviously, you can't hold the weapon system itself accountable," Docherty says. It would also be legally challenging and arguably unfair to hold the operator responsible for the actions of a system that was functioning autonomously, she adds.

Russell suggests that there be "no communication between the on-board computing and the firing circuit". That means the firing has to be activated by a remote operator and cannot ever be activated by the AI.

There is at least one point in the LAWs discussions that (almost) everybody seems to agree on: even nations generally opposed to controls, including the United States and China, have indicated that autonomous agents, including those with AI, should play no part in the decision to launch nuclear weapons, says Russell.

However, Russia seems to be more circumspect on this issue. Moscow is widely thought to have resurrected a cold-war programme called Perimetr, which – in theory at least – could launch a first nuclear strike on the West with no human oversight[6]. The United States and China have raised this issue in various talks about autonomous weapons, which many say could put pressure on Russia to change its strategy.

Policing the system

Unfortunately, says Kallenborn, any ban on the use of LAWs would be hard to enforce through inspections and observations – the classic 'trust but verify' approach commonly used for other regulated weaponry.

With nuclear weapons, for example, there's a well-established system for site inspections and audits of nuclear material. But with AI, things are easier to conceal or alter on the fly. "It could be as simple as just changing a couple lines of code to say, all right, now the machine gets to decide to go blow this up. Or, you know, remove the code, and then stick it back in when the arms-control inspectors are there," says Kallenborn. "It requires us to rethink how we think about verification in weapons systems and arms control."

Checks might have to switch from time-of-production to after-the-fact, Kallenborn says. "These things are going to get shot down. They're going to be captured. Which means that you can then do inspections and look at the code," he says.

All these issues will feed into the UN discussions, beginning at the General Assembly this September. If enough countries vote to act in September, then the UN will probably set up a working group to set out the issues, Docherty says.

A treaty might be possible in three years, adds Docherty, who had a key role in the negotiations of the UN's 2017 Treaty on the Prohibition of Nuclear Weapons. "In my experience, once negotiations start, they move relatively quickly."

David Adam is a writer in Hertford, near London.

1. Russell, S. *Nature* **614**, 620–623 (2023).
2. Taddeo, M. & Blanchard, A. *Sci. Eng. Ethics* **28**, 37 (2022).
3. Bode, I. & Watts, T. *Loitering Munitions and Unpredictability* (Center for War Studies, Univ. Southern Denmark, 2023).
4. David, S. Preprint at bioRxiv https://doi.org/10.1101/2023.09.03.556099 (2023).
5. Su, J., Vargas, D. V. & Sakurai, K. *IEEE Trans. Evol. Comput.* **23**, 828–841 (2019).
6. Topychkanov, P. in *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Volume I Euro-Atlantic Perspectives* (ed. Boulanin, V.) Ch. 8 (SIPRI, 2019).