Comment

YASUYOSHI CHIBA/AFP VIA GETTY

A Russian-launched, Iranian Shahed-136 loitering missile flies over Kyiv in October 2022.

AI weapons: Russia's war in Ukraine shows why the world must enact a ban

Stuart Russell

Conflict pressures are pushing the world closer to autonomous weapons that can kill without human control. Researchers and the international community must join forces to prohibit them.

One year since Russia's invasion, an arms race in artificial-intelligence (AI) weaponry is being played out on Ukrainian soil. Western audiences cheer when plucky Ukrainian forces use modified commercial quadcopters to drop grenades on Russian soldiers. They boo when brutal Russian forces send swarms of cheap Iranian cruise missiles to destroy hospitals, power plants and apartment blocks. But this simple 'us versus them' narrative obscures a disturbing trend — weapons are becoming ever smarter.

Soon, fully autonomous lethal weapon systems could become commonplace in conflict. Some are already on the market. Mercifully, few have actually been used in warfare, and none has been used in Ukraine, at the time of writing. Yet evolving events are a cause for concern.

The inevitable logic of using electronic countermeasures against remotely operated weapons is driving both sides towards increasing the level of autonomy of those weapons. That is pushing us ever closer to a dangerous world where lethal autonomous weapon systems are cheap and widely available tools for inflicting mass casualties — weapons of mass destruction found in every arms supermarket, for sale to any dictator, warlord or terrorist.

Although it is difficult to discuss banning weapons that might help the Ukrainian cause, it is now urgent that world governments do so and limit the use of AI in war. No one wants this bleak future of robotic threats.

As a start, governments need to begin serious negotiations on a treaty to ban anti-personnel autonomous weapons, at the

very least. Professional societies in AI and robotics should develop and enforce codes of conduct outlawing work on lethal autonomous weapons. And people the world over should understand that allowing algorithms to decide to kill humans is a terrible idea.

Pressures leading to full autonomy

What exactly are 'lethal autonomous weapons systems'? According to the United Nations, they are "weapons that locate, select, and engage human targets without human supervision". The word 'engage' in this definition is a euphemism for 'kill'. I am not talking about weapons that are operated remotely by humans, such as the US Predator drone or Ukraine's home-made grenade droppers, because these are not autonomous. Nor am I talking about anti-missile defence systems, or about the fully autonomous drones that both Russians and Ukrainians are using for reconnaissance, which are not lethal. And I am not talking about the science-fiction robots portrayed in the 'Terminator' films – controlled by the spooky emergent consciousness of the Skynet software system and driven by hatred of humanity – that the media often conjure up when discussing autonomous weapons. The issue here is not rogue machines taking over the world, but weapons deployed by humans that will drastically reduce our physical security.

Current AI systems exhibit all the required capabilities – planning missions, navigating, 3D mapping, recognizing targets, flying through cities and buildings, and coordinating attacks. Lots of platforms are available. These include: quadcopters ranging from centimetres to metres in size; fixed-wing aircraft (from hobby-sized package-delivery planes and full-sized, missile-carrying drones to 'autonomy-ready' supersonic fighters, such as the BAE Systems Taranis); self-driving trucks and tanks; autonomous speedboats, destroyers and submarines; and even skeletal humanoid robots.

The road to full autonomy in the Russia–Ukraine conflict begins with various types of semi-autonomous weapon already in use. For example, Russia is deploying 'smart' cruise missiles to harsh effect, hitting predefined targets such as administrative buildings and energy installations. These weapons include Iranian Shahed missiles, nicknamed 'mopeds' by the Ukrainians owing to their sound, which can fly low along rivers to avoid detection and circle an area while they await instructions. Key to these attacks is the use of swarms of missiles to overwhelm air-defence systems, along with

JIM HUYLEBROEK/*NEW YORK TIMES*/REDUX/EYEVINE

Ukrainian soldiers operate a surveillance drone on the front line near Kherson, Ukraine.

minimal radio links to avoid detection. I have heard reports that new Shaheds are being equipped with infrared detectors that enable them to home in on nearby heat sources without requiring target updates communicated from controllers by radio – if true, this would be an important step towards full autonomy.

The Ukrainians have deployed Turkish Bayraktar teleoperated weapons against tanks and other targets since the early days of the war. Improved Russian air defences and jamming have made these weapons more vulnerable and less effective over time; moreover, they cost around US$5 million each (250 times more expensive than Shaheds). Commercial, remote-controlled quadcopters that have been adapted to drop grenades have proved effective in small-scale tactical operations, and remotely piloted boats have been used to attack naval targets. But, as jamming systems become the norm, teleoperation becomes more difficult and autonomous weapons increasingly attractive.

> **"Human life would be devalued if robots take life-or-death decisions."**

Elsewhere, lethal autonomous weapons have been on sale for several years. For example, since 2017, a government-owned manufacturer in Turkey (STM) has been selling the Kargu drone, which is the size of a dinner plate and carries 1 kilogram of explosive. According to the company's website in 2019 (since toned down), the drone is capable of "autonomous and precise" hits against vehicles and persons, with "targets selected on images" and by "tracking moving targets" (see go.nature.com/3ktq6bb). As reported by the UN, Kargu drones were used in 2020 by the Libyan Government of National Accord – despite a strict arms embargo – to autonomously 'hunt down' retreating forces[1].

Other 'loitering' forms of missile, such as the Shahed, also exhibit a form of autonomy. The Israeli Harpy drone can fly over a region for several hours looking for targets that match a visual or radar signature and then destroy them with its 23-kilogram explosive payload. (Russia's Lancet missile, widely used in Ukraine, has similar characteristics.) Whereas the Kargu and Harpy are 'kamikaze' weapons, the Chinese Ziyan Blowfish A3 is an autonomous helicopter equipped with a machine gun and several unguided gravity bombs. All of these systems are described as having both autonomous and remotely operated modes, making it difficult to know whether any given attack was carried out by a human operator.

Benefits and problems

Why are militaries pursuing machines that can decide for themselves whether to kill humans? Like remotely operated weapons, autonomous aircraft, tanks and submarines can carry out

missions that would be suicidal for people. They are cheaper, faster, more manoeuvrable and have longer range than their crewed counterparts; can withstand higher *g*-forces in flight; and function underwater without life-support systems. But, unlike remotely operated weapons, autonomous weapons can function even when electronic communication is impossible because of jamming — and can react even faster than any weapon remotely controlled by a human. AI expert Kai-Fu Lee, among others, has described autonomous weapons as the 'third revolution in warfare' after gunpowder and nuclear weapons[2].

A common argument in favour is that waging war through autonomous weapons will protect military lives, just as remotely operated weapons and cruise missiles are said to do. But this is a fallacy. The other side would have such weapons, too — and as we have seen in Ukraine, the death toll among soldiers as well as civilians is staggering.

Another point often advanced is that, compared with other modes of warfare, the ability of lethal autonomous weapons to distinguish civilians from combatants might reduce collateral damage. The United States, along with Russia, has been citing this supposed benefit with the effect of blocking multilateral negotiations at the Convention on Certain Conventional Weapons (CCW) in Geneva, Switzerland — talks that have occurred sporadically since 2014.

The case relies on two claims. First, that AI systems are less likely to make mistakes than are humans — a dubious proposition now, although it could eventually become true. And second, that autonomous weapons will be used in essentially the same scenarios as human-controlled weapons such as rifles, tanks and Predator drones. This seems unequivocally false. If autonomous weapons are used more often, by different parties with varying goals and in less clear-cut settings, such as insurrections, repression, civil wars and terrorism, then any putative advantage in distinguishing civilians from soldiers is irrelevant. For this reason, I think the emphasis on the weapons' claimed superiority in distinguishing civilians from combatants, which originates from a 2013 UN report[3] pointing to the risks of misidentification, has been misguided.

There are many more reasons why developing lethal autonomous weapons is a bad idea. The biggest, as I wrote in *Nature* in 2015 (ref. 4), is that "one can expect platforms deployed in the millions, the agility and lethality of which will leave humans utterly defenceless". The reasoning is illustrated in a 2017 YouTube video advocating arms control, which I released with the Future of Life Institute (see go.nature.com/4ju4zj2). It shows 'Slaughterbots' — swarms of cheap microdrones using AI and facial recognition to assassinate political opponents. Because no human supervision is required, one person can launch an almost unlimited number of weapons to wipe out entire populations. Weapons experts concur that anti-personnel swarms should be classified as weapons of mass destruction (see go.nature.com/3yqjx9h). The AI community is almost unanimous in opposing autonomous weapons for this reason.

Moreover, AI systems might be hacked, or accidents could escalate conflict or lower the threshold for wars. And human life would be devalued if robots take life-or-death decisions, raising moral and justice concerns. In March 2019, UN secretary-general António Guterres summed up this case to autonomous-weapons negotiators in Geneva: "Machines with the power and discretion to take lives without human involvement are politically unacceptable, morally repugnant and should be prohibited by international law" (see go.nature.com/3yn6pqt). Yet there are still no rules, beyond international humanitarian laws, against manufacturing and selling lethal autonomous weapons of mass destruction.

"So far, no academic society has developed a policy on autonomous weapons."

Political action at a standstill

Unfortunately, politics has not kept up with technological advances. Dozens of human-rights and arms-control organizations have joined the Campaign to Stop Killer Robots, which calls for a ban on lethal autonomous weapons. Politicians and governments have failed to act, despite polls suggesting broad public support for such a ban (more than 60% of adults; see, for example, go.nature.com/416myef). Thousands of researchers and leaders in AI, including me, have joined those calls (see go.nature.com/4gqmfm5), yet, so far, no academic society has developed a policy on autonomous weapons because of concerns about discussing matters that are not purely scientific.

One reason that negotiations under the CCW have made little progress is confusion, real or feigned, about technical issues. Countries still argue endlessly about the meaning of the word 'autonomous'. Absurdly, for example, Germany declared that a weapon is autonomous only if it has "the ability to learn and develop self-awareness". China, which ostensibly supports a ban on autonomous weapons, says that as soon as weapons become capable of autonomously distinguishing between civilians and soldiers, they no longer count as autonomous and so wouldn't be banned. The United Kingdom has pledged never to develop or use lethal autonomous weapons, but keeps redefining them so that its pledge is effectively meaningless. For example, in 2011, the UK Ministry of Defence wrote that "a degree of autonomous operation is probably achievable now", but in 2017 stated that "an autonomous system is capable of understanding higher-level intent". Michael Fallon, then secretary of state for defence, wrote in 2016 that "fully autonomous systems do not yet exist and are not likely to do so for many years, if at all", and concluded that "it is too soon to ban something we simply cannot define" (see go.nature.com/3xrztn6).

Further progress in Geneva soon is unlikely. The United States and Russia refuse to allow negotiations on a legally binding agreement. The United States worries that a treaty would be unverifiable, leading other parties to circumvent a ban and creating a risk of strategic

MEHMET KAMAN/ANADOLU AGENCY VIA GETTY

A fleet of Kargu drones at Turkish manufacturer STM in Ankara.

ED RAM/*GUARDIAN*/EYEVINE

The aftermath of a 'kamikaze' drone attack in Kyiv.

surprise. Russia now objects that it is being discriminated against, because of its invasion of Ukraine.

A pragmatic way forward

Rather than blocking negotiations, it would be better for the United States and others to focus on devising practical measures to build confidence in adherence. These could include inspection agreements, design constraints that deter conversion to full autonomy, rules requiring industrial suppliers to check the bona fides of customers , and so on. It would make sense to discuss the remit of an AI version of the Organization for the Prohibition of Chemical Weapons, which has devised similar technical measures to implement the Chemical Weapons Convention. These measures have neither overburdened the chemical industry nor curtailed chemistry research. Similarly, the New START treaty between the United States and Russia allows 18 on-site inspections of nuclear-weapons facilities each year. And the Comprehensive Nuclear-Test-Ban Treaty might never have come into existence, had not scientists from all sides worked together to develop the International Monitoring System that detects clandestine violations.

Despite the impasse in Geneva, there are glimmers of hope. Of those countries that have stated a position, the vast majority favours a ban. Negotiations could progress in the UN General Assembly in New York City, where no country has a veto, and at ministerial-level meetings. Last week, the government of the Netherlands hosted a meeting in The Hague on 'responsible AI in the military domain', where the question of whether it is ethical to introduce this class of weapon at all was raised. During the meeting, the United States announced a "political declaration" of principles and best practices for the military use of AI and urged other nations to sign up to these (see go.nature.com/3xsj779). Perhaps the most important is the statement that: "States should maintain human control and involvement for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment." Already, more than 60 countries, including China, have joined the declaration. Unfortunately, it is non-binding and does not rule out any category of autonomous weapon.

On 23–24 February, Costa Rica is due to host a meeting of Latin American and Caribbean nations on the 'social and humanitarian impact of autonomous weapons', which includes threats from non-state actors who might use them indiscriminately. These same nations organized the first nuclear-weapon-free zone, raising hopes that they might also initiate a treaty declaring an autonomous-weapon-free zone.

Next steps

In my view — and I suspect that of most people on Earth — the best solution is simply to ban lethal autonomous weapons, perhaps through a process initiated by the UN General Assembly. Another possibility, suggested as a compromise measure by a group of experts (see go.nature.com/3jugzxy) and formally proposed to the international community by the International Committee of the Red Cross (see go.nature.com/3k3tpan), would ban anti-personnel autonomous weapons. Like the St Petersburg Declaration of 1868, which prohibited exploding ordnance lighter than 400 grams, such a treaty could place lower limits on the size and payload of weapons, making it impossible to deploy vast swarms of small devices that function as weapons of mass destruction.

Instead of blocking progress in Geneva, countries should engage with the scientific community to develop the technical and legal measures that could make a ban on autonomous weapons verifiable and enforceable. Technical questions include the following. What physical parameters should be used to define the lower limit for allowable weapons? What are 'precursor' platforms (which can be scaled up to full autonomy), and how should their production and sale be managed? Should design constraints be used, such as requiring a 'recall' signal? Can firing circuits be separated physically from on-board computation, to prevent human-piloted weapons from being converted easily into autonomous weapons? Can verifiable protocols be designed to prevent accidental escalation of hostilities between autonomous systems?

On the civilian side, professional societies in AI and robotics (including the Association for the Advancement of Artificial Intelligence, the Association for Computing Machinery and the Institute of Electrical and Electronics Engineers) should develop and enforce codes of conduct proscribing work on lethal autonomous weapons. There are many precedents: for example, the American Chemical Society has a strict chemical-weapons policy (see go.nature.com/3yn8ajt) and the American Physical Society asks the United States to ratify the Comprehensive Nuclear-Test-Ban Treaty (see go.nature.com/3jrajvr), opposes the use of nuclear weapons against non-nuclear states (see go.nature.com/3k4akq8) and advocates robust research programmes in verification science and technology for the benefit of peace and security (see go.nature.com/3hzjkkv).

As Russia's war in Ukraine unfolds, and as autonomous-weapons technology races ahead (along with the desire to use it), the world cannot afford another decade of diplomatic posturing and confusion. Governments need to deliver on what seems a simple request: to give their citizens some protection against being hunted down and killed by robots.

The author

Stuart Russell is professor of computer science, cognitive science, and computational precision health at the University of California, Berkeley, California, USA.
e-mail: russell@berkeley.edu

1. United Nations. *Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011)* (UN Security Council, 2021).
2. Lee, K.-F. 'The third revolution in warfare' *The Atlantic* (11 September 2021).
3. UN General Assembly. *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns* (United Nations, 2013).
4. Russell, S. *Nature* **521**, 415–418 (2015).

The author declares no competing interests.