FOREIGN AFFAIRS

FEBRUARY 29, 2024

The Perilous Coming Age of AI Warfare

How to Limit the Threat of Autonomous Weapons

PAUL SCHARRE

The Perilous Coming Age of AI Warfare

How to Limit the Threat of Autonomous Weapons

PAUL SCHARRE

Last year, the Ukrainian drone company Saker claimed it had fielded a fully autonomous weapon, the Saker Scout, which uses artificial intelligence to make its own decisions about who to kill on the battlefield. The drone, Saker officials declared, had carried out autonomous attacks on a small scale. Although this has not been independently verified, the technology necessary to create such a weapon certainly exists. It is a small technical step—but a consequential moral, legal, and ethical one—to then produce fully autonomous weapons that are capable of searching out and selecting targets on their own.

The deployment of Saker's drone shows that the window to regulate autonomous weapons is closing fast. Countries have been discussing what to do about autonomous weapons for a decade, but they have been unable to agree on regulations to limit the weapons' harms. Yet there is an urgent need for international agreement. The unconstrained development of autonomous weapons could lead to wars that expand beyond human control, with fewer protections for both combatants and civilians. Even if a wholesale ban is not realistic, there are many practical regulations that governments can adopt to mitigate autonomous weapons' worst dangers. Without limits, humanity risks barreling toward a future of dangerous, machine-driven warfare.

NEARLY THERE

Militaries have used partially autonomous weapons in limited, defensive circumstances since the 1980s. Today, at least 30 countries operate air and missile defense systems, or antirocket protection systems for ground vehicles, that have autonomous modes. Once activated, these defensive systems can automatically sense incoming rockets, artillery, mortars, missiles, or aircraft and intercept them. But humans supervise their operation and can intervene if something goes wrong.

Militaries have been slow to adopt AI technology developed in the commercial sector, in part because of cumbersome procurement processes. The war in Ukraine accelerated innovation on both sides, especially with commercial technologies such as small drones. Both Moscow and Kyiv have used drones extensively for reconnaissance and attacks on ground forces. These drone deployments have in turn led to the development of new countermeasures, including electronic warfare systems that jam drones' communications links or pinpoint the location of the operators on the ground, who can then be attacked. This focus on the operator is strategically sound: most drones are remotely piloted, making the human operator essential—and a critical target. Without human operators, remotely controlled drones become useless. That is why autonomous drones are so valuable: they do not rely on vulnerable communications links. To counter them, the drones themselves must be found and destroyed.

The specific form that autonomous weapons take will depend on the needs of a conflict. In Ukraine, Moscow and Kyiv have used small aerial drones to target personnel and attack vehicles. Larger, medium-altitude drones have been used to reach deeper behind enemy lines to target radars and installations. Ukraine has even used drone boats to attack the Russian Black Sea Fleet. All these drones, which are currently remotely controlled, could be upgraded to become autonomous, allowing continued operation if the communications link were jammed.

Other conflicts could lead to the development of different autonomous weapons. Several countries—including China, France, India, Russia, the United Kingdom, and the United States—are currently working on stealth combat drones. Future wars could see these drones autonomously targeting air defenses or mobile missile launchers. Ground robots have lagged behind their air and sea counterparts, but future wars could see autonomous weapons deployed on robots or fixed gun emplacements. The changes will likely not stop there. Swarms of drones could autonomously coordinate their behavior, reacting to changes on the battlefield at a speed beyond human capabilities. Autonomous reactions at machine speed could drive a faster tempo of operations, accelerating the pace of battle. This in turn could create even more pressure to eliminate humans from decision cycles. The consequences of this shift to a new era of machine-driven warfare will be profound.

AUTONOMOUS PERIL

Leading AI scientists, including the University of California professor Stuart Russell and the Turing Award winner Yann LeCun, have warned of the dangers of autonomous weapons. A consortium of over 250 nongovernmental organizations, including Amnesty International, Human Rights Watch, and the Nobel Women's Initiative, have formed the Campaign to Stop Killer Robots, calling for a preemptive, legally binding treaty to ban autonomous weapons. These warnings and campaigns are motivated by concern that autonomous weapons could increase civilian casualties in war. Although autonomous weapons could conceivably reduce civilian casualties by precisely targeting combatants, in the hands of a state that cares little about civilian casualties—or wants to punish a civilian population—they could be used to commit devastating atrocities. Massive hordes of autonomous weapons could be deployed to target and kill thousands at a time, making today's smart bombs seem clumsy by comparison.

One of the most extreme risks comes from integrating AI and autonomy into nuclear weapons. In 2022, the United States declared that it would always retain a "human 'in the loop'" for decisions to use nuclear weapons. The United Kingdom adopted a similar policy in 2022. Yet Russia and China have not. Human control over nuclear weapons seems like an easy starting point for international agreement, but Moscow has shown a disturbing willingness to integrate risky automation into its nuclear operations. This is nothing new: after the Cold War ended, former Soviet officials explained that the Soviet Union had built a semiautomated retaliatory nuclear strike system called "Perimeter." Once activated, it would use a series of automated sensors to detect a nuclear attack on Soviet soil. If one was detected and there was no response from the country's leaders—presumably because they had been killed in the attack—the system would automatically transfer nuclear launch authority to a relatively junior officer in a secure bunker. Russian officials stated in 2018 that the system is still operational and has even been upgraded. More recently, Moscow has begun to develop a nuclear-armed autonomous underwater drone. Nuclear-armed drones at sea or in the air could be sent on patrol, risking accidents or losing control of a nuclear weapon.

Widely deployed autonomous weapons integrated with other aspects of military AI could result in a new era of machine-driven warfare. Military AI applications can accelerate information processing and decision-making. Decision cycles will shorten as countries adopt AI and automation to reduce the time to find, identify, and strike enemy targets. In theory, this could allow for more time for humans to make thoughtful, deliberate decisions. In practice, competitors will feel forced to respond in kind, using automation to speed up their own operations to keep pace. The result will be an escalating spiral of greater automation and less human control.

The end state of this competition will likely be war executed at machine speed and beyond human control. In finance, the widespread use of algorithms in high-frequency trading has led to stocks being traded autonomously at superhuman speeds. The Chinese military scholar Chen

Hanghui of the People's Liberation Army's Army Command College has hypothesized about a "singularity" on the battlefield, a point wherein the pace of machine-driven warfare will similarly outstrip the speed of human decision-making. This tipping point would force humans to cede control to machines for both tactical decisions and operational-level war strategies. Machines would not only select individual targets but also plan and execute whole campaigns. The role of humans would be reduced to switching on the machines and sitting on the sidelines, with little ability to control or even end wars.

STUCK IN THE WEEDS

International regulations, if carefully crafted and successfully implemented, could help mitigate some of the worst harms of autonomous weapons. Around 30 countries and a consortium of humanitarian organizations have called for a preemptive, legally binding treaty to ban autonomous weapons before they can be deployed. Governments have had relative success banning chemical and biological weapons, cluster munitions, and the use of the environment as a weapon. But similar progress on regulating autonomous weapons has proved challenging. An all-out ban, in particular, is unlikely. Because autonomous weapons have not yet been fully developed, both their potential harms and their military value are unknown. Governments are, therefore, reluctant to give up a potentially valuable weapon because of uncertain claims about potential future harms.

Discussions about how to regulate autonomous weapons are taking place in many forums. On an international level, governments have been discussing autonomous weapons at the UN Convention on Certain Conventional Weapons since 2014. The CCW is an international forum for regulating weapons that are deemed to be excessively harmful to combatants or civilians, such as land mines and blinding lasers. It includes 126 countries, and agreement requires the support of all participating governments, which is a recipe for dysfunction. Russia and the United States, in particular, have staunchly opposed a treaty banning autonomous weapons, arguing that existing rules in the law of war are sufficient to address any potential harms. Moscow's and Washington's opposition is fatal, as a ban on autonomous weapons is meaningless if it does not include the world's major military powers. Recognizing the lack of progress, in 2023, proponents of a ban took the issue to the UN General Assembly, which does not require consensus. The First Committee of the General Assembly voted in November 2023 to task the UN secretary-general with preparing a report on autonomous weapons, which advocates of a ban hope will be the first step toward a mandate for negotiating a treaty.

The United States, meanwhile, has proposed an alternative approach to an outright ban, and in late 2023, it led over 40 countries in endorsing a political declaration on the need for the responsible use of military AI. As of February 2024, over 50 governments had joined the effort. Although the declaration does not ban autonomous weapons, it does provide general guidelines for using them, such as ensuring adequate testing to reduce the risk of accidents. Although it espouses valuable principles, the declaration lacks meaningful restrictions on the most dangerous forms of autonomous weapons, such as antipersonnel autonomous weapons and autonomy in nuclear operations.

BEFORE IT IS TOO LATE

Instead of these stalled approaches, there are five complementary initiatives that countries could pursue to reduce the threat of autonomous weapons. First, governments could adopt a broad principle that establishes the minimum necessary human involvement in lethal decision-making. Such a principle could be adopted as a legally binding rule, either through the CCW or the General Assembly, or it could be a politically binding declaration. This standard should require that a human decision-maker has specific and sufficient information about the intended target, the weapon, the environment, and the context for the attack to determine whether an attack is lawful before it can be authorized. Countries could also agree that in order for human control to be meaningful, any use of autonomous weapons must be limited in geography, time, and the targets being attacked. Although such a principle would not ban all autonomous weapons—in fact, it would legitimize their use when following these rules—it would provide guardrails around how countries use autonomous weapons and ensure human involvement in the most critical decisions for authorizing attacks.

Second, governments could ban autonomous weapons that target people. In a world in which many combatants do not wear uniforms, even humans often struggle to accurately distinguish civilians from soldiers. Algorithms would have a much harder time concluding whether a person holding a rifle is a combatant or a farmer protecting his land. A machine is also less likely to be able to accurately and reliably recognize whether a soldier is genuinely trying to surrender or is only feigning surrender. As a consequence, antipersonnel autonomous weapons pose greater risks than those that target only vehicles or equipment, and the harm that these weapons could inflict far exceeds their military value. States could choose to ban them before their use becomes widespread.

Third, countries could promulgate best practices for testing military AI and autonomous systems to avoid accidents. A key aim of the 2023 U.S.-led political declaration is to ensure that military AI systems are safe by improving testing. Washington should go further, sharing with other countries the best practices for testing AI systems to improve their safety. This could be done in a way similar to how the United States shares information on its process for conducting legal weapons reviews, without releasing the content of specific reviews.

States could choose to ban antipersonnel autonomous weapons before their use becomes widespread.

Fourth, the United States should partner with the United Kingdom to persuade other nuclear powers to pursue an agreement ensuring strict human control over nuclear weapons. London is the right partner for this effort because of its policy of maintaining human control over its nuclear arsenal. Securing similar unilateral statements from other nuclear powers or, ideally, a multilateral agreement will be an important step toward ensuring that humanity's most dangerous weapons remain in the control of humans. An agreement among the five permanent members of the UN Security Council would be a powerful statement. Washington should also press Beijing on this issue in the new U.S.-Chinese bilateral AI talks.

Finally, states could adopt uniform rules of the road for autonomous drones to reduce the risk of accidents. As countries field more drones in the air and at sea—and as these drones become increasingly autonomous—the chance that an accident or miscalculation could trigger an international incident becomes more likely. In 2019, for example, Iranian air defenses shot down a U.S. Global Hawk drone over the Strait of Hormuz. In March 2023, a Russian fighter jet interfered with a U.S. MQ-9 Reaper drone in the Black Sea, causing the drone to crash. Because these drones were controlled remotely, a human decided how to respond. But future air and sea drones could be autonomous. This would mean that if similar situations were to occur, the drone would autonomously take the actions it had been programmed to do. If it had been programmed to fire back, it would do so, potentially escalating an international incident without any deliberate human decision to do so. Governments must ensure that any autonomous behaviors are consistent with human intent. The 1972 U.S.-Soviet Incidents at Sea Agreement helped reduce the number of unplanned incidents between the U.S. and Soviet navies during the Cold War. A similar autonomous incidents agreement would help countries navigate the risks of dueling autonomous systems deployed in contested areas and avoid unplanned and unwanted incidents.

These initiatives to deal with the risks of autonomous weapons could be pursued together or separately, through different channels and over different timelines. The CCW, the General Assembly, the U.S.-led political declaration initiative, the Security Council, U.S.-Chinese bilateral dialogue, and other multilateral forums are all spaces for global cooperation that should be considered. Washington's collaboration with Beijing will be particularly important. Agreement between the United States and China, as the world's leading military, economic, and technological superpowers, will be a difficult yet critical step forward in managing the risks of autonomous weapons.

WATCH OUT

Autonomous weapons are coming. Attempts to ban them entirely, although well meaning, are likely futile. Their military value is simply too great. Yet countries have a choice about how autonomous weapons will be used. Without effective restrictions, autonomous weapons will reduce human control over warfare, pose increased danger to civilians and combatants, and undermine international stability. Steps must be urgently taken to address these weapons' worst dangers. Doing so will require moving beyond the current simplistic and misleading choice between a ban on all autonomous weapons and no restrictions at all.

Autonomous weapons are an early test of humanity's ability to deal with weaponized AI, more dangerous forms of which are coming. Cutting-edge AI systems have demonstrated the ability to aid in the development of cyberweapons and chemical and biological weapons. Global cooperation is urgently needed to govern their improvement, limit their proliferation, and guard against their potential use. Reaching international agreement on autonomous weapons is critical for addressing their harms and laying the foundation for collaboration on future, even more consequential AI dangers.