The US and 30 Other Nations Agree to Set Guardrails for Military AI

The tech-centric war in Ukraine and the success of ChatGPT have prompted new interest in figuring out how to prevent military AI from going awry.

By Will Knight

Nov 08, 2023 01:37 PM · 5 min. read · View original

When politicians, tech executives, and researchers gathered in the UK last week to discuss the risks of artificial intelligence, one prominent worry was that algorithms might someday turn against their human masters. More quietly, the group made progress on controlling the use of AI for military ends.

Content

To honor your privacy preferences, this content can only be viewed on the site it originates from.

On November 1, at the US embassy in London, US vice president Kamala Harris announced a

range of AI initiatives, and her warnings about the threat AI poses to human rights and democratic values got people's attention. But she also revealed a declaration signed by 31 nations to set guardrails around military use of AI. It pledges signatories to use legal reviews and training to ensure military AI stays within international laws, develop the technology cautiously and transparently, avoid unintended biases in systems that use AI, and continue to discuss how the technology can be developed and deployed responsibly.

"A principled approach to the military use of AI should include careful consideration of risks and benefits, and it should also minimize unintended bias and accidents," the declaration says. It also says that states should build safeguards into military AI systems, such as the ability to disengage or deactivate when a system demonstrates "unintended behavior."

The declaration is not legally binding, but it is the first major agreement between nations to impose voluntary guardrails on military AI. On the same day, the UN announced a new resolution from its General Assembly that calls for an in-depth study of lethal autonomous weapons and could set the terms for restrictions on such weapons.

Lauren Kahn, a senior research analyst at the Center for Security and Emerging Technology

(CSET) at Georgetown University, calls the US-led declaration “incredibly significant.” She says it could offer a practical path toward binding international agreement on the norms around how nations develop, test, and deploy AI in military systems, providing greater safeguards and transparency around applications involving weapons systems. “I really believe that these are common sense agreements that everyone would agree to,” Kahn says.

The nonbinding declaration was first drafted by the US following a conference attended by representatives from different nations that focused on military use of AI and was held in The Hague in February. The US has also asked other nations to agree that humans remain in control of nuclear weapons. The new declaration states that the nations behind it will meet in early 2024 to continue discussions.

Vice President Harris announced during her speech in London that the declaration has now been signed by US-aligned nations that include the UK, Canada, Australia, Germany, and France. The 31 signatories do not include China or Russia, which alongside the US are seen as leaders in the development of autonomous weapons systems. China did join with the US in signing a declaration on the risks posed by AI as part of the AI Safety Summit coordinated by the British government.

Deadly Automation

Talk of military AI often evokes the idea of AI-powered weapons capable of deciding for themselves when and how to use lethal force. The US and several other nations have resisted calls for an outright ban on such weapons, but the Pentagon's policy is that autonomous systems should allow "commanders and operators to exercise appropriate levels of human judgment over the use of force." Discussions around the issue as part of the UN's Convention on Certain Conventional Weapons—established in 1980 to create international rules around the use of weapons deemed to be excessive or indiscriminate in nature—have largely stalled.

The US-led declaration announced last week doesn't go so far as to seek a ban on any specific use of AI on the battlefield. Instead, it focuses on ensuring that AI is used in ways that guarantee transparency and reliability. That's important, Kahn says, because militaries are looking to harness AI in a multitude of ways. Even if restricted and closely supervised, the technology could still have destabilizing or dangerous effects.

One concern is that a malfunctioning AI system might do something that triggers an escalation in hostilities. "The focus on lethal autonomous weapons is important," Kahn says. "At the same

time, the process has been bogged down in these debates, which are focused exclusively on a type of system that doesn't exist yet."

Some people are still working on trying to ban lethal autonomous weapons. On the same day Harris announced the new declaration on military AI, the First Committee of the UN General Assembly, a group of nations that works on disarmament and weapons proliferation, approved a new resolution on lethal autonomous weapons.

The resolution calls for a report on the "humanitarian, legal, security, technological, and ethical" challenges raised by lethal autonomous weapons and for input from international and regional organizations, the International Committee of the Red Cross, civil society, the scientific community, and industry. A statement issued by the UN quoted Egypt's representative as saying "an algorithm must not be in full control of decisions that involve killing or harming humans," following the vote.

"It's an exciting, momentous time," says Anna Hehir, program manager for autonomous weapons systems at the Future of Life Institute, a nonprofit that campaigns for an outright ban on lethal autonomous systems that target humans. "It's a big step toward getting to a legally binding instrument, which the UN

Secretary General has called for to happen by 2026."

Boom Times

Militaries around the world have long been interested in AI, but the rapid deployment of new technologies on the battlefield in Ukraine has prompted renewed interest from the US and others. The Pentagon is experimenting with incorporating AI into smaller, cheaper systems as a way to increase its capacity to sense threats and react rapidly.

"The systems that we're starting to see play out in Ukraine are unprecedented—it's technology that we haven't seen before," Hehir says of the widespread use of drones in the conflict, including some with AI for identifying targets. "It's definitely a playground for testing out different technologies."

ChatGPT doesn't appear to have been conscripted into military service yet, but the recent flourishing of chatbot technology seems to have prompted renewed and more serious debate around the risks of military AI. "The political declaration and the UN vote signify a pretty significant change in the debate surrounding autonomous weapons for the last several years," says Paul Scharre, an expert on autonomous weapons and director of studies at the Center for New American Security (CNAS), a think tank in Washington, DC.

Some autonomous weapons already exist, including defensive systems aboard battleships that can automatically shoot down incoming missiles. But there have only been a couple of reports of potential use of lethal systems that incorporate modern AI in warfare. A drone deployed during the civil war in Libya in 2020 by forces backed by the government in Tripoli may have used lethal force against soldiers without human control, according to a 2021 UN report. There are also some reports of lethal autonomous drones being developed for Ukrainian forces trying to repel Russia's renewed invasion. Russia is among the nations that dissented from the new UN resolution, saying the agreement would undermine existing work on autonomy under the Convention on Certain Conventional Weapons.