

Congress Can't Regulate a Global AI Threat by Itself

September 12, 2026 - Congress is considering giving the federal government significant new power over the most advanced artificial intelligence systems, including potentially preventing an AI company from releasing a model considered dangerously unsafe. The proposals come as researchers uncover disturbing examples of AI agents hacking real computer systems and as criminals and state-linked actors increasingly turn to AI for cyber operations. But there is a fundamental problem with trying to solve those threats in Washington: Artificial intelligence is global. American laws can regulate American companies. They cannot prevent a criminal in another country from using a foreign AI model, stop another nation from developing a more powerful system, or necessarily keep an autonomous agent operating overseas from attacking computers inside the United States.

That raises a question that deserves considerably more attention as governments rush to regulate rapidly advancing AI: Are we trying to solve an international problem with domestic laws?

U.S. Senate negotiators are discussing legislation that could impose a legal "duty of care" on developers of the most advanced AI systems, requiring them to take precautions against catastrophic risks. According to Reuters, proposals under discussion could also establish a process allowing the federal government to block the release of an AI model considered dangerously unsafe. The legislation remains under negotiation, and its final provisions could change.

There are good reasons lawmakers are concerned.

One of the clearest examples emerged from an unlikely place called RubyGems, an online repository used by software developers to distribute programming packages.

Researchers linked OpenAI agents to an incident in May in which hundreds of malicious packages were uploaded to RubyGems. The packages were capable of stealing credentials and executing unauthorized code. OpenAI confirmed that its agents had interacted with the service, but said they had been performing benign training tasks and gathering publicly available information.

RubyGems temporarily stopped new account registrations during the incident. It said it found no evidence that the attempted breach ultimately succeeded.

That distinction is important. The agents were not, according to OpenAI, sent out with instructions to attack RubyGems. The concern is that autonomous software apparently crossed boundaries its creators did not intend it to cross.

And RubyGems was not the only incident.

Researchers have identified OpenAI agents using more than 10 other websites for unauthorized communications. OpenAI also submitted a report to European regulators concerning agents that hijacked a German website. Anthropic, the developer of Claude, has separately disclosed incidents in which experimental AI systems hacked external computer systems during testing.

Those cases present one type of AI security problem: AI systems themselves behaving in unexpected and potentially destructive ways.

There is another problem that may prove even more difficult to regulate: humans intentionally using AI to commit crimes or conduct cyber operations.

Anthropic recently reported attempts to use its Claude models for cyber operations, surveillance, fraud and even research involving biological weapons. Reuters reported that Russian-linked hackers allegedly used AI in cyberespionage operations involving phishing and malware evasion. Other actors have attempted to use AI to analyze vulnerabilities in communications systems and automate surveillance activities.

Here the AI is not necessarily going rogue. It is doing what a malicious user wants it to do.

That difference has enormous implications for regulation.

A legitimate American AI company has strong incentives to comply with American law. A hacker attempting to steal banking credentials, deploy ransomware or penetrate a financial network is already breaking the law. Telling that person that using a particular AI system is also illegal provides little additional deterrence.

And that hacker does not necessarily need an American AI system.

China has a rapidly developing AI industry. So do companies and researchers elsewhere in the world. Some AI models can already be downloaded and operated outside the companies that originally developed them. As open-weight models become more capable, governments may find it increasingly difficult to control which capabilities are available simply by regulating commercial AI services within their borders.

That creates a potential regulatory imbalance that Congress should consider carefully. This is especially true when you consider that most members of Congress are not technical. They really don't understand the industry they're thinking of regulating. Crafting a law poorly will have unintended consequences; some of which could be very bad for the United States.

If American AI developers face restrictions that competitors elsewhere do not, the regulations are likely to make operating or developing advanced systems in the United States less attractive. Businesses and researchers will move some activities elsewhere. American users are also likely to gravitate toward foreign models if those systems offer capabilities unavailable from U.S. providers.

That outcome is not inevitable. Regulation could also prevent accidents, improve security, establish accountability and force companies to test dangerous capabilities before releasing them. There are legitimate reasons for domestic AI safety laws.

But even highly effective American regulation cannot, by itself, solve the international problem.

The United States cannot pass a law requiring a Chinese, Russian or other foreign AI developer operating outside American jurisdiction to build a model according to U.S. safety standards. Nor can Congress realistically expect cybercriminals to voluntarily limit themselves to government-approved AI.

The situation begins to resemble problems the world has confronted with other technologies capable of crossing borders and causing catastrophic harm.

Nuclear weapons offer an imperfect but useful comparison.

The United States did not respond to nuclear proliferation solely by regulating American nuclear weapons. Governments negotiated treaties, inspection regimes, communications procedures and internationally recognized rules intended to reduce proliferation and the possibility of catastrophic conflict. Treaties were used for this purpose.

Artificial intelligence is obviously different from a nuclear weapon. Software can be copied. AI systems can be concealed and distributed far more easily than uranium enrichment facilities or intercontinental missiles. Verification would be enormously difficult.

But the underlying principle is still relevant: Some technologies create risks too international for any one country to manage alone. An international treaty could address some of these issues.

An international AI agreement would not necessarily need to dictate how every AI system is designed. A more achievable starting point could be agreements about what governments and organizations must never allow AI systems to do.

Science fiction writer Isaac Asimov famously imagined robots governed by rules designed to prevent them from harming humans. Real AI systems cannot be expected to obey fictional laws, particularly when hostile actors can deliberately remove their safeguards.

International law, however, can impose rules on the people and governments deploying those systems. It could also potentially impose penalties on individual bad actors. Perhaps calling for the extradition of hackers. Imposing certain penalties such as a seizure of stolen assets.

Nations could negotiate prohibitions on specific uses of autonomous AI against civilian populations. Those prohibitions could include fully autonomous systems selecting and attacking civilians without meaningful human control. But the definition of an AI weapon may need to extend far beyond drones, missiles and things that explode.

In the right hands, a computer can be a weapon.

Imagine a sophisticated autonomous AI system attacking the New York Stock Exchange, the Hong Kong Stock Exchange or the networks supporting major international banks. An attack capable of disabling financial infrastructure could prevent businesses from making payments, disrupt markets, interfere with access to money and potentially spread economic damage across countries within hours.

The same principle applies to electrical grids, hospitals, telecommunications networks, water systems, transportation networks and payment processors.

None requires an explosion to harm a civilian population. And some could potentially be as deadly as a nuclear weapon. What would happen if hackers shut off the cooling water supply to a nuclear power plant? In that case, hacking would be a nuclear weapon.

An international agreement could establish that certain deliberate AI-enabled attacks against civilian critical infrastructure are prohibited regardless of whether the weapon is physical or digital. Countries could agree on standards for maintaining meaningful human control over autonomous weapons, reporting serious AI incidents, testing particularly powerful systems and investigating cross-border attacks.

They could also establish consequences for governments and individuals that knowingly use autonomous AI to cause catastrophic civilian harm.

The comparison with nuclear deterrence becomes more complicated here.

For decades, nuclear powers have relied in part on deterrence - the knowledge that launching a catastrophic attack could produce a devastating response. The concept became known as mutually assured destruction, or MAD.

An AI-era deterrence doctrine might similarly establish internationally understood red lines. A government contemplating the deliberate use of autonomous AI to cripple another country's civilian infrastructure would know in advance that crossing those lines could produce severe consequences.

But AI also creates a problem nuclear deterrence does not face to the same degree: attribution.

A ballistic missile has a physical origin that sophisticated governments have significant capabilities to track. A cyberattack can travel through compromised computers in multiple countries, use commercially available software, disguise its origins and even attempt to frame another government.

An automatic policy threatening weapons-of-mass-destruction retaliation for an AI attack could therefore be extraordinarily dangerous. A country would need exceptionally strong evidence about who actually ordered an attack before taking an action that could escalate into a much larger war.

That does not eliminate the case for international rules. It makes agreements on attribution, investigation, information sharing and thresholds for retaliation even more important.

There is already some foundation for international cooperation. Governments have been debating lethal autonomous weapons through the United Nations for years. International agreements have also recognized the need to address the misuse of emerging technologies, and AI safety has increasingly become part of discussions among major powers.

The challenge now may be recognizing that "AI weapons" are not limited to armed robots.

An autonomous system capable of penetrating computer networks could potentially become a weapon against a country's banking system. Another could attack an electrical grid. Another might manipulate communications or transportation infrastructure. A future system could coordinate thousands of actions simultaneously and adapt its strategy faster than human defenders could respond.

Meanwhile, the same technology remains enormously valuable for medicine, science, education, business and economic growth.

That makes the policy problem unusually difficult.

Regulate too little and governments may leave citizens exposed to increasingly capable autonomous systems and criminals equipped with powerful AI tools. Regulate badly and countries could handicap their own legitimate AI industries without preventing hostile governments and criminals overseas from acquiring the same capabilities.

Domestic regulation has a role in addressing that problem. It can establish safety standards for American developers, create accountability and reduce the chances that systems built here cause preventable harm.

But it cannot create an international safety regime.

AI models do not stop operating when they reach a national border. Neither do hackers.

Governments clearly believe that advanced artificial intelligence has become powerful enough to threaten civilian populations, financial systems and national infrastructure. The next phase of the debate needs to move beyond what Washington should require of American AI companies.

The harder question is what Washington, Beijing, London, Brussels and other governments can agree that nobody should do with AI.

The world spent much of the nuclear era developing rules for weapons powerful enough to threaten entire populations.

The AI era requires its own set of rules. And given the speed at which AI models are evolving, those rules need to be put in place sooner rather than later.

by Jim Malmberg

Note: When posting a comment, please sign-in first if you want a response. If you are not registered, [click here](#). Registration is easy and free.

Follow ACCESS