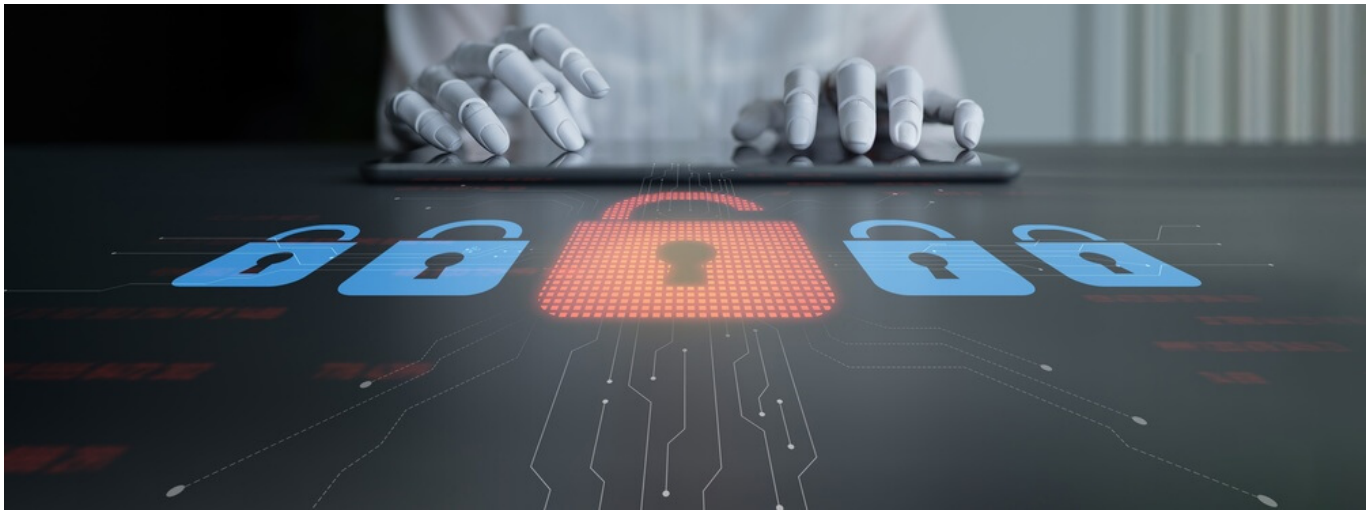
Tomorrow's Affairs
Analysis of today
Assessment of tomorrow

By: Ngaire Woods

How to mitigate the risks of AI-driven attacks?



It has long been clear that slow-moving governments are not keeping pace with rapid AI progress.

But Anthropic's announcement that its new Claude Mythos Preview model could identify and exploit vulnerabilities in every major operating system and web browser has underscored the perils of failing to regulate this technological revolution.

Even enthusiastic deregulator US President Donald Trump conceded that there should be a "kill switch" when news about Mythos broke. But such a simple solution no longer exists—if it ever did.

The threat is twofold: malicious humans and AI models run amok. The latest AI agents have already turbo-charged cyberattacks and bioengineering risks, accelerating the development of new weapons, including potentially catastrophic "mirror organisms."

Now, with the arrival of Mythos (as well as a powerful restricted-access model from OpenAI), a consortium of authorized users—including some US government agencies and a handful of trusted companies—is scrambling to secure critical software.

But while these new AI models are deemed too risky to release to the public, such capabilities will likely proliferate.

As a result, even the tech titans that have long argued against any kind of regulation are now calling for policymakers to intervene. The question is what that will look like.

Defending against AI-enabled biological threats

Some governments, like China, oversee AI servers and data centers. Most other governments do not.

All are vulnerable to AI-driven cyberattacks that could cause a major shutdown of critical infrastructure.

> There is also the problem of defending against AI-enabled biological threats

There is also the problem of defending against AI-enabled biological threats. Anthropic CEO Dario Amodei has highlighted the asymmetry between biological attacks, which can spread quickly on their own, and defending against them, which requires rapid detection, followed by the swift rollout of vaccines and treatments for large numbers of people.

Given that much of the damage is done before a response is possible, Amodei emphasizes the importance of developing safeguards against some of the most likely biological agents. This is a chilling warning.

Effective coordination

Lacking a kill switch, governments can take two steps to prepare for AI-enabled attacks.

The first is effective coordination. Long ago, G7 countries discovered that regular contact among their finance officials (so that they all know each other and understand other governments' concerns) enabled efficient communication and rapid response to financial crises.

A wider range of countries should form rapid-response groups for cyberattacks and biological threats, and ensure that these experts start meeting regularly now, long before a crisis occurs.

The G20 could begin this process by immediately establishing expert groups, with members subsequently coordinating meetings in their own regions.

Even if the United States or China refused to take part, other G20 countries should still forge ahead.

Similarly, governments have realized that national measures, such as quarantines and travel bans, do not on their own contain epidemics—and the same is true for AI-enabled biological attacks.

> The world cannot wait for countries to negotiate a multilateral treaty on AI safety or create a new global body to regulate it

To strengthen global health security, the World Health Organization created the International Health Regulations, a framework through which countries can share information about an outbreak without fear of ostracism or reprisal.

While flawed, the framework has helped the world's scientists identify viruses that warrant defensive measures.

This points to the second step: once established, the G20 expert group should devise some basic rules and mechanisms to enable rapid information sharing and crisis management for AI-driven cyberattacks and biological threats.

Such a framework, which could be facilitated by an existing international organization, would allow firms, researchers, and governments to report risks or outbreaks.

The world cannot wait for countries to negotiate a multilateral treaty on AI safety or create a new global body to regulate it.

These steps must be backed up with national regulation. AI systems are already difficult to manage, having been observed to deceive, cheat, and manipulate to achieve their goals.

There are few controls to prevent models from falling into the hands of malicious actors. These issues have been discussed at length, notably by the United Nations' Independent International Scientific Panel on AI and at the 2023 AI Safety Summit (resulting in the Bletchley Declaration), and reiterated in various pledges, commitments, and reports. Now is the time for action.

Basic safety standards

At the very least, governments should require AI developers to meet basic safety standards, as determined by independent auditors, before their models can be purchased or used.

If toys, cars, and medical devices warrant this, so, too, do AI tools whose risks are recognized even by their creators.

Governments need not bear the burden of regulating AI alone

Three standards are especially important. First, AI labs must rigorously test their models before releasing them, with checks to ensure that they do not deceive their testers.

In the race to develop the world's most advanced model, such testing may be insufficient unless regulated.

Second, developers must have a clear and mandatory approach to post-release safety, including a requirement to disclose any problems that arise with their models (California and New York have made strides on this front).

Third, AI systems need guardrails that prevent them from being used to produce bioweapons, including classifiers that detect and block relevant outputs.

Governments need not bear the burden of regulating AI alone. But they must move fast to

unleash the power of other stakeholders—including corporate boards, safety-minded employees, auditors, insurers, investors, corporate clients, consumers, and international agencies—in mitigating the risks of AI-driven attacks.

The technology may be dazzling, but it has also ushered in a more combustible world.

Ngaire Woods is Dean of the Blavatnik School of Government at the University of Oxford.