National Cyber Security Centre | Te Tira Tiaki Government Communications Security Bureau

The implications of frontier AI models

Cyber Security Advisory

*The Traffic Light Protocol (TLP) marking is used to ensure that sensitive information is shared with the correct audience. Information sources may use **TLP: CLEAR** when information carries minimal or no foreseeable risk of misuse, in accordance with applicable rules and procedures for public release.*

Advice from the National Cyber Security Centre

Background

There has been significant commentary in recent weeks about the cyber security implications that Frontier AI models such as Mythos could mean for the cyber security community. The NCSC has developed some high-level advice to support senior leaders and network defenders to consider the implications of frontier AI products.

About Frontier AI

Frontier AI represents the most advanced models of AI software. Frontier AI models have demonstrated the ability to discover vulnerabilities in software products. Malicious actors can use these newly found vulnerabilities to exploit systems at a greater speed and scale than before.

A recent report from Anthropic about their product Mythos Preview raises questions about the implications of frontier AI models for cyber security. As an agentic model, Mythos Preview can autonomously complete a series of tasks. Anthropic says it can identify zero-day vulnerabilities in code and then weaponise them into fully working exploits.

Other frontier models can also identify vulnerabilities in code, although creating a working exploit is not an automated process, and AI guardrails can make it harder to do.

As new vulnerabilities continue to be discovered, the best line of defence remains effective security controls. NCSC recommends that organisations review their current security posture to ensure that it remains fit for purpose and that appropriate methods to detect and contain malicious activity are implemented across the network.

Senior leaders should consider the following:

Senior leaders should consider having a conversation with their cyber security team and seek regular reporting about how potential system vulnerabilities are identified and managed.

The conversation can be informed through questions such as:

- Do we have a vulnerability management programme, and does it need to change in response to the potential proliferation of vulnerabilities which could be identified through frontier AI products?

- How well does our vulnerability management programme currently operate?
- How would it manage if we had to increase the frequency of our patching operations?
- Do we have a vulnerability disclosure policy?
- If you use software developed in-house, what processes do you have to identify and fix vulnerabilities quickly?
- What processes do you have to quickly identify and fix vulnerabilities or programmes developed in house?
- How can we get assurance from our third-party suppliers to ensure we have assurance on their systems?
- What protections do we have to ensure that any suspicious activity can be detected and contained?
- What plans do we currently have to respond to incidents, and what were the results of our last tabletop exercise?
- How are we implementing security controls for our critical systems?
- How could better resourcing improve our security posture?

Network defenders should consider the following:

- Developers should consider how to safely involve frontier models in code reviews. They could look for vulnerabilities, including in open-source dependencies, before a software update
- Patch frequently, prioritising systems exposed to the internet.
- Reduce the attack surface, and apply defence in depth principles to prevent a breach from progressing. For example:
    - Minimise exposure of systems and services to the internet
    - Segment networks
    - Uninstall unused applications
    - Disable unused services/accounts
    - Use multi-factor authentication
    - Filter out malicious traffic
    - Review the vulnerability management policies of your software and system supply chains including whether and how AI is being used for finding vulnerabilities.
    - Frequently monitor for potential compromise, investigating suspicious behaviour on both the network and on endpoints.
    - Review the Minimum Cyber Security Standards.
    - Read the New Zealand Information Security Manual (NZISM) for additional guidelines specific to your organisation.

The NCSC can be contacted by email at: **info@ncsc.govt.nz**. We encourage you to contact us at any time if you require any further assistance or advice.