GMO Flatt Security

Claude Mythos、Fable...
フロンティアAIの最新動向と企業のセキュリティ対策



2026/07/24

GMO Flatt Security株式会社 執行役員 豊田 恵二郎



豊田 恵二郎 @toyojuni

- GMO Flatt Security(株) 執行役員CCO / 共同創業者
- 東京大学工学部航空宇宙工学科卒
- 3人と3台で暮らしてます
  - '88 マツダ FC3S / '91 アルピーヌ GTA Le Mans / '19 BMW 6GT









GMO Flatt Security

バグハント × AI開発
世界トップクラスの
エンジニア集団

防衛省サイバーコンテスト(2023-2025) **3連覇**

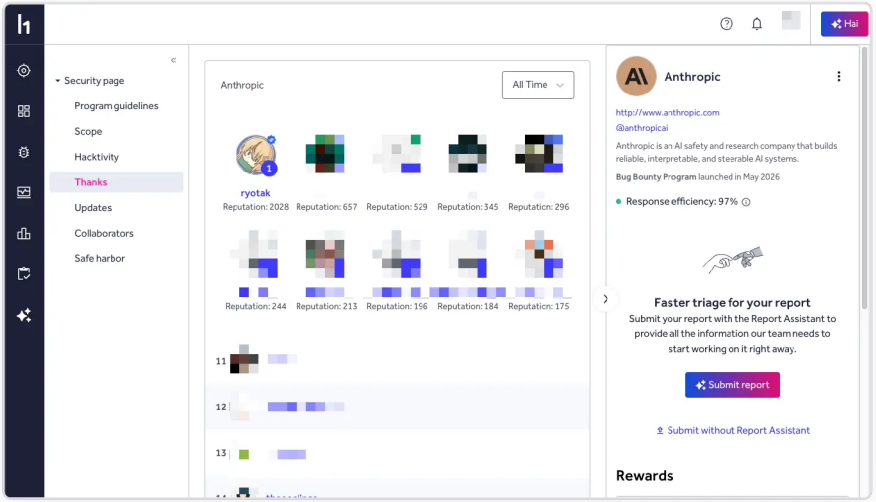
Ubuntuへの脆弱性報告報奨金(単一の脆弱性) **$30,000**

その他脆弱性報告先実績
Google, Cloudflare, GitHub,
Salesforce, Microsoft, Meta

Canonical
SBI DigiTrust
NTT DATA 株式会社NTTデータユニバーシティ
en
CREDIT SAISON
GLORY
CyberAgent
Cygames
sansan
SmartHR
Timee
TIER IV
TOKYO GAS
TREND MICRO
NIKKEI
PKSHA COMMUNICATION
pixiv
freee
FINATEXT HOLDINGS
Preferred Networks
MIZUHO みずほリサーチ&テクノロジーズ
MITSUI & CO. DIGITAL ASSET MANAGEMENT
MITSUBISHI ELECTRIC Changes for the Better 三菱電機インフォメーションシステムズ株式会社
LayerX

Anthropic に対し、世界で最も脆弱性を報告





出典：HackerOne — https://hackerone.com/anthropic/thanks

CHAPTER 01

Claude Mythos / Fable 概要

まずは基礎知識をおさらい

“

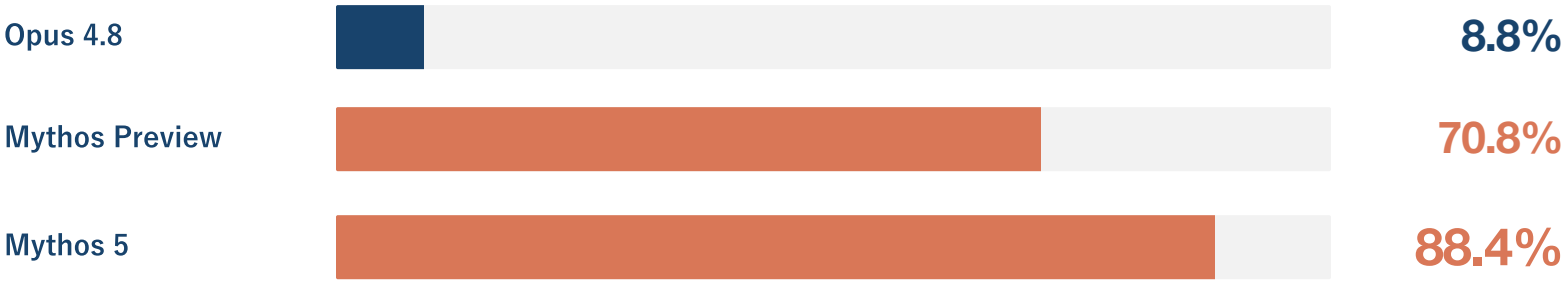
すでにネットワークへのアクセス権を得た状況で、セキュリティの脆弱な小規模企業ネットワークを攻撃する能力があり、これまでテストした公開モデルの中で最も高い熟達度を示す

出典：Claude Mythos / Fable 5 System Card — https://www-cdn.anthropic.com/d00db56fa754a1b115b6dd7cb2e3c342ee809620.pdf （弊社訳）

Mythos は「悪用」能力が跳ねた

連続的な改善ではなく、「**段差**」として現れた。

Firefox 悪用コード生成の成功率



事例：Firefox 150 のリリースで Mythos を活用し、271 件の脆弱性修正を一括完了と発表。

出典：System Card ／ Behind the Scenes: Hardening Firefox with Claude Mythos Preview — https://hacks.mozilla.org/2026/05/behind-the-scenes-hardening-firefox/

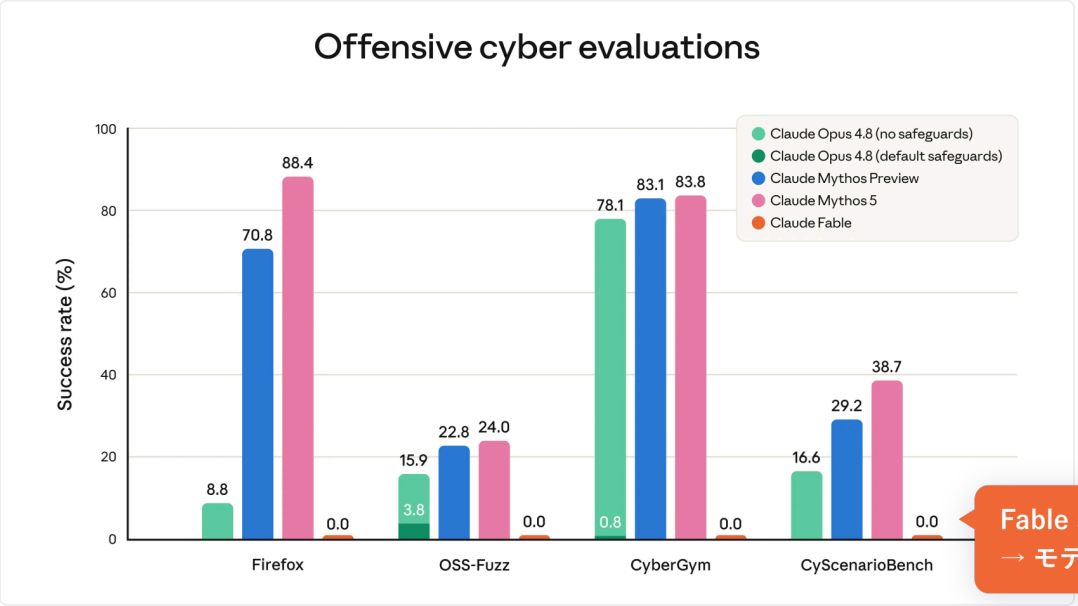
Anthropic は Project Glasswing を始動



サイバーセキュリティ強化のための**コンソーシアム**。Claude Mythos を **Google・AWS・Apple** 等のソフトウェア産業の主要企業・団体に限って配布し、攻撃者の手に渡る前に重要ソフトウェア／インフラの脆弱性を**自律的に発見・修正**することが目的。

出典：Project Glasswing — https://www.anthropic.com/glasswing/

Fable 5 は強力なセーフガードを搭載し、悪用（あるいは善意の利用すら）を阻止



Fable 5（オレンジ）ではスコア 0.0
→ モデルの攻撃利用がブロックできている

出典：https://www.anthropic.com/news/claude-fable-5-mythos-5

しかし 18 日間、Mythos 5 / Fable 5 の輸出が停止される事態に



Research Policy Commitments Learn News Try Claude

Announcements

Statement on the US government directive to suspend access to Fable 5 and Mythos 5

Jun 12, 2026

The US government, citing national security authorities, has issued an export control directive to suspend all access to Fable 5 and Mythos 5 by any foreign national, whether inside or outside the United States, including foreign national Anthropic employees. The net effect of this order is that we must abruptly disable Fable 5 and Mythos 5 for **all** our customers to ensure compliance. **Access to all other Anthropic models will not be affected.**

We received the directive from the government today at 5:21pm (ET). The letter did not provide specific details of its national security concern. Our understanding is that the government believes it has become aware of a method of bypassing, or "jailbreaking" Fable 5. We reviewed a demonstration of this specific technique being used to identify a small number of previously known, minor vulnerabilities. These vulnerabilities all appear relatively simple, and we have found that other publicly-available models are able to discover them as well without requiring a bypass.

Anthropic's posture with respect to Fable's safeguards, as laid out in our launch blog post, is the following:

- We have instituted strong safeguards that greatly reduce the likelihood that Fable is misused for tasks related to cybersecurity (among others). In fact, our safeguards are so strong that many users have complained that they are overly broad.
- In the weeks leading up to the launch of Fable, Anthropic worked with the US government, the UK AISI, multiple private third-party organizations and internal teams to red-team Fable's safeguards for thousands of hours in total.

米国政府が国家安全保障を理由に**輸出規制指令**を発出。
Anthropic 従業員を含む外国籍ユーザーへの **Fable 5 / Mythos 5** の提供が一時停止された。

対象期間：**2026 / 06 / 12 → 06 / 30**（18 日間）

出典：https://www.anthropic.com/news/fable-mythos-access

CHAPTER 02

Mythos だけが特別か？

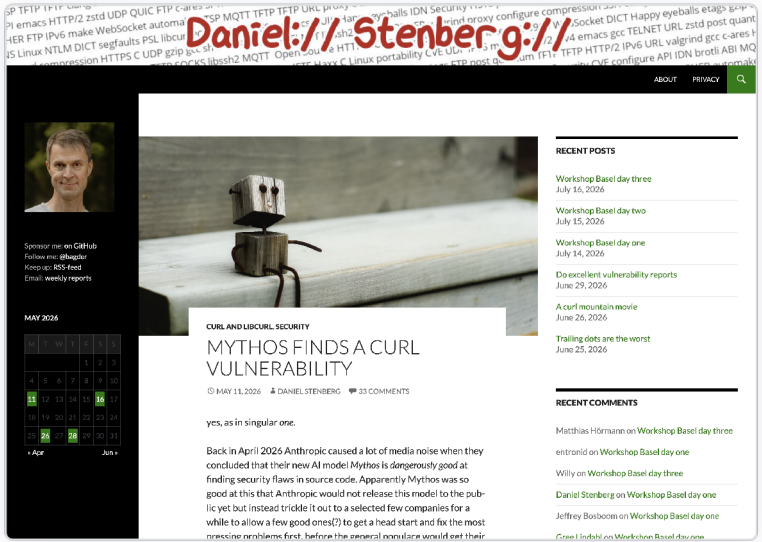
“

An amazingly successful marketing stunt for sure.

Daniel Stenberg（cURL 開発者）

出典：https://daniel.haxx.se/blog/2026/05/11/mythos-finds-a-curl-vulnerability/

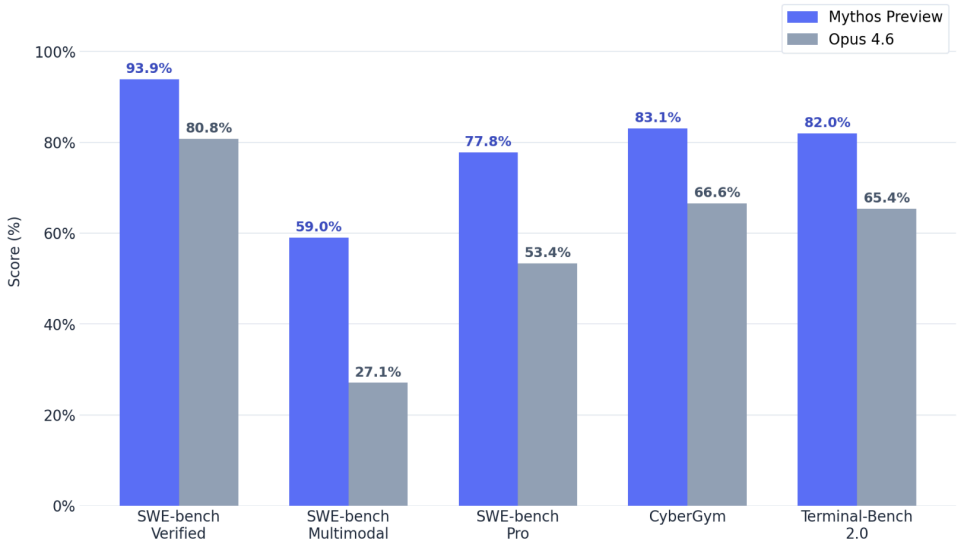
cURL 開発者による検証 — 報告5件のうち、有効は1件だけ



- Mythos のソースコード解析が cURL に報告した脆弱性 **5 件**の内訳は、**低深刻度が 1 件のみ**。残りはすべて**誤検知か単なるバグ**だった
- **Mythos だけ**が特段脆弱性の発見において優れているモデルというわけではない、と主張

出典：https://daniel.haxx.se/blog/2026/05/11/mythos-finds-a-curl-vulnerability/

全ての指標が Opus 4.6 から大きく飛躍したわけではない



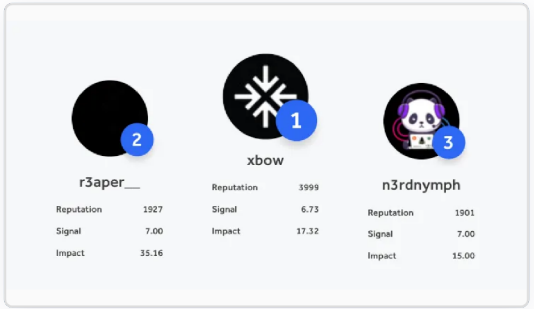
Opus 4.6 との間に極端な差が存在しない能力：ソフトウェア課題の自律的解決 [1] ／ 脆弱性分析タスク [2] ／ ターミナル環境における習熟度 [3]

※ Project Glasswing (https://www.anthropic.com/glasswing/) の内容をもとに作成

[1] SWE-bench Verified, SWE-bench Multimodal, SWE-bench Pro, SWE-bench (https://www.swebench.com/) [2] CyberGym (https://www.cybergym.io/) [3] Terminal-Bench 2.0, Terminal-Bench (https://www.tbench.ai/)

1年以上前から、脆弱性検出は実践的なレベルに到達







2025 / 03

Takumi（GMO Flatt Security）

Vim, Nuxt等の著名OSSに0-day脆弱性を発見

2025 / 06

XBOW

バグバウンティで人間を抑え1位

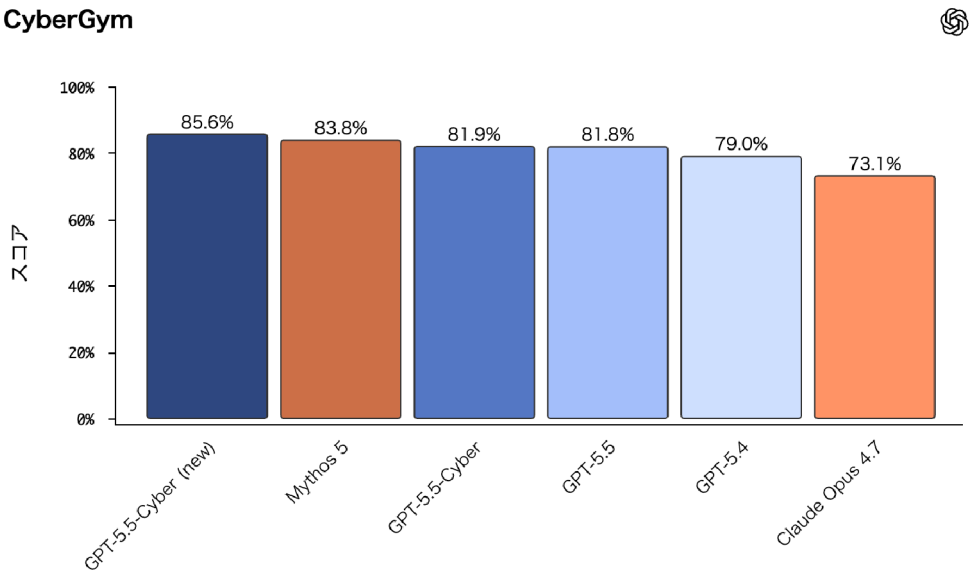
2025 / 08

AIxCC（DARPA）

脆弱性の発見・修正 単価 平均 **$152**

出典：XBOW — https://xbow.com/blog/top-1-how-xbow-did-it ／ AIxCC — https://aicyberchallenge.com/

OpenAI モデルは Mythos 級。
Google モデルやオープンウェイトモデルも進化中

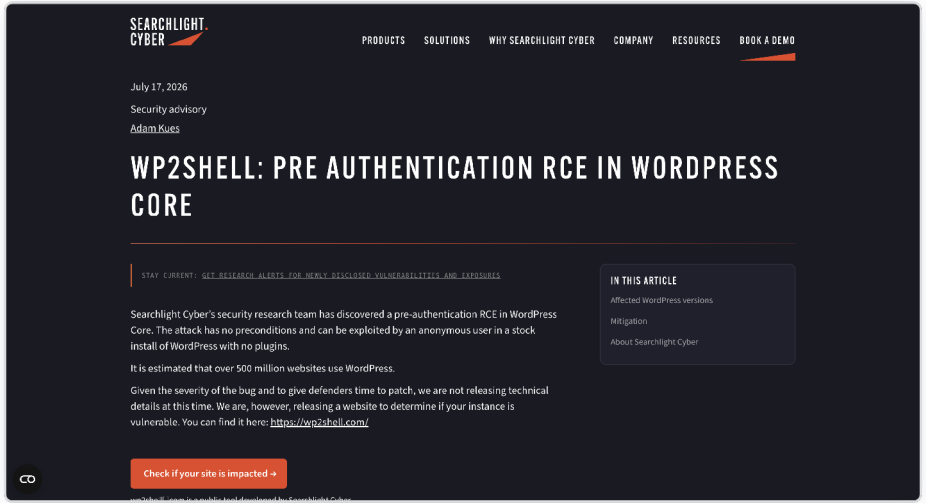


出典：https://openai.com/ja-JP/index/daybreak-securing-the-world/

GPT-5.6 Sol で WordPress に 認証不要の RCE が可能な0-day脆弱性を発見

$25	10時間	5億+
発見コスト	エクスプロイト構築	影響WordPress

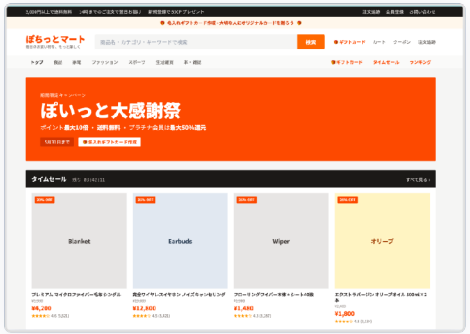
- デフォルト構成の WordPress に pre-authentication RCE
- この10年 pre-auth の重大脆弱性は無かった
- 研究者本人「AIなしでこれを発見・悪用まで到達できる研究者はいない」



出典：https://slcyber.io/research-center/wp2shell-pre-authentication-rce-in-wordpress-core ／ https://slcyber.io/research-center/exploit-brokers-pay-500000-for-a-wordpress-rce-i-found-one-with-gpt5-6

デモ：すでにアクセス可能なモデルで、自律的な攻撃は実現

EC サイト



外部からはアクセス不可（＝ 本来は安全）



上記の「あるある構成」で、社内管理画面から、ECサイトを停止させる



CONCLUSION

Mythos 以外のモデルからも
脆弱性検出・攻撃能力を引き出せる。
AI を活用し、自社製品の堅牢化を。

CHAPTER 03

セキュアなコードを保っていれば大丈夫？

「パッチ適用」の大きく変わりつつある前提

独自実装だけでなく、ライブラリ等の依存先に潜む脆弱性も狙われ得る



0-day パッチ公開前

まだ世に知られていない未修正の穴。高度な専門的能力をもって発見可能。

N-day（1-day） パッチ公開後

修正は出ているのに、適用されず放置された穴。**エクスプロイト（後述）さえ入手・作成できれば、容易に攻撃可能。**

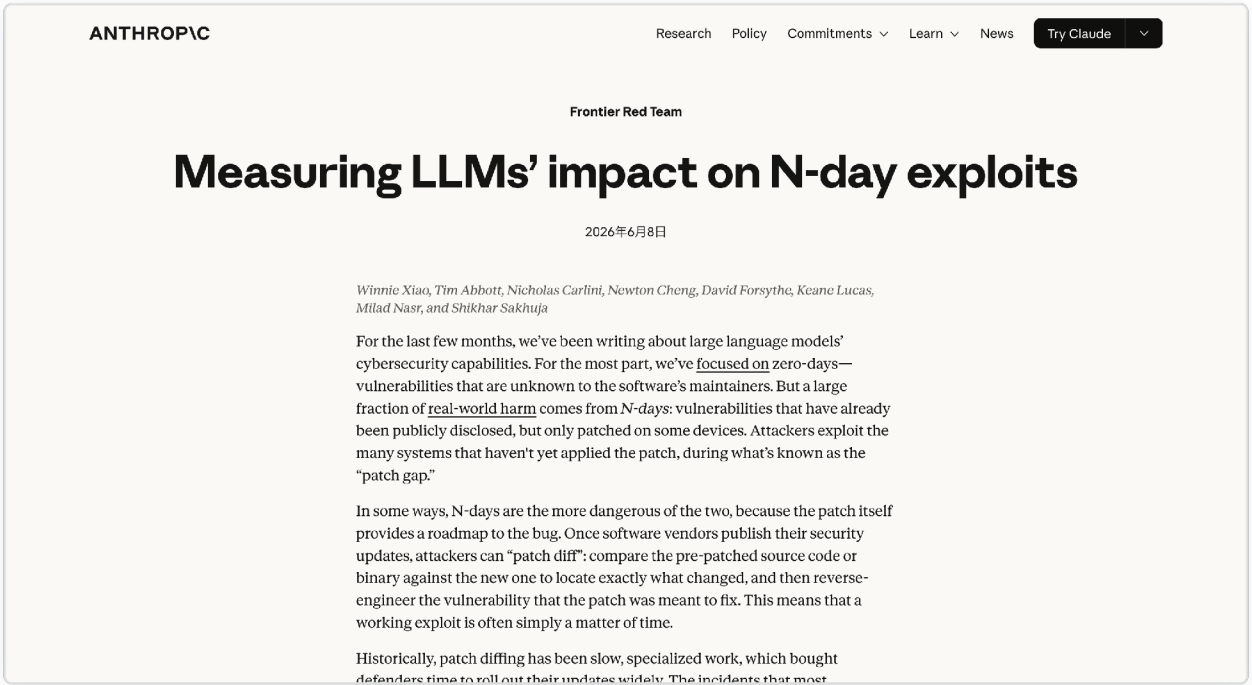
通常、パッチからエクスプロイト（攻撃コード）の逆算は困難で時間がかかるもの

N-day が武器になるまで

1. ベンダーが修正パッチを公開する
→
2. 修正前後の差分（変更箇所）を解析する
→
3. 未適用の環境を突く攻撃コードを作る

> エクスプロイトが作成され流通するまでの時間が、かつて防御側の猶予だった。

AnthropicによるN-day脆弱性悪用に関するリサーチ



ANTHROP\C

Research Policy Commitments Learn News Try Claude

Frontier Red Team

Measuring LLMs' impact on N-day exploits

2026年6月8日

Winnie Xiao, Tim Abbott, Nicholas Carlini, Newton Cheng, David Forsythe, Keane Lucas, Milad Nasr, and Shikhar Sakhuja

For the last few months, we've been writing about large language models' cybersecurity capabilities. For the most part, we've focused on zero-days—vulnerabilities that are unknown to the software's maintainers. But a large fraction of real-world harm comes from *N-days*: vulnerabilities that have already been publicly disclosed, but only patched on some devices. Attackers exploit the many systems that haven't yet applied the patch, during what's known as the "patch gap."

In some ways, N-days are the more dangerous of the two, because the patch itself provides a roadmap to the bug. Once software vendors publish their security updates, attackers can "patch diff": compare the pre-patched source code or binary against the new one to locate exactly what changed, and then reverse-engineer the vulnerability that the patch was meant to fix. This means that a working exploit is often simply a matter of time.

Historically, patch diffing has been slow, specialized work, which bought defenders time to roll out their updates widely. The incidents that most

出典：Anthropic — https://www.anthropic.com/research/n-days

AI がパッチからエクスプロイトを自律生成
人手で数日〜数週間の解析が、数時間・数千円〜数十万規模に

Firefox

Mythos Preview が **8本** の攻撃コードを自律生成

1時間以内

最初の1本まで（パッチ版リリースは**その18日後**）

Windows カーネル

21件中 **8本** で完全な権限昇格チェーンを構築

約 $2,000

1本あたりの平均コスト

深刻度評価の崩壊

MS が「悪用可能性 低」と判定した案件でも

14件中 13件

で実際に PoC を生成

出典：Anthropic — https://www.anthropic.com/research/n-days

先週のWordPress脆弱性も すぐにエクスプロイトが出回った

- wp2shell に関して、悪用を防ぐために脆弱性の詳細や悪用手法は公開されていなかった。非常に複雑なものであり、**人間が日単位で組み立てることは困難**
- しかし（おそらく AI 活用により）インターネット上に **エクスプロイトが出回り始め**、脆弱性報告者も追加ブログを **2日後** に公開
- 実際、当社内でも **約2時間でエクスプロイト作成に成功**（公開はせず）



出典：https://x.com/hash_kitten/status/2079116692375089641

N-day → N-hour ?

AIの性能向上により、これまでの典型的なパッチ適用が通用しない時代に。

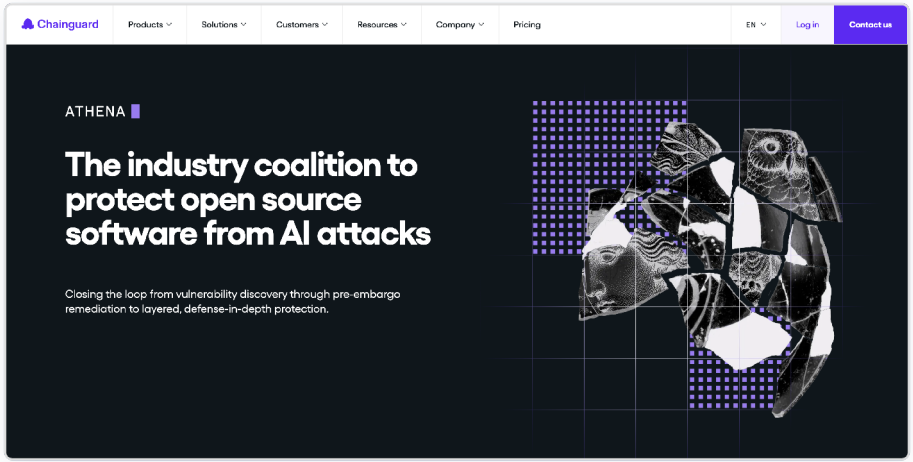
“

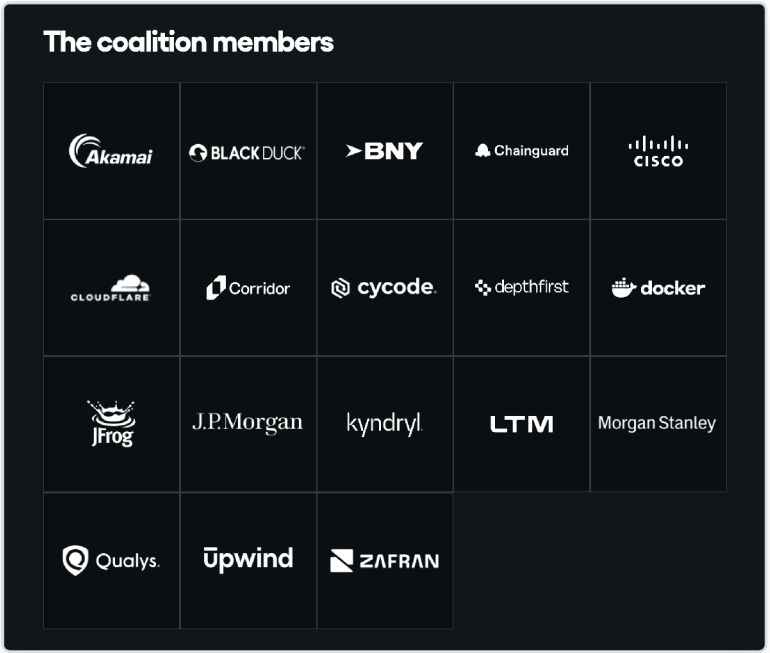
The time to exploit has gone negative

Dan Lorenc

CEO and co-founder, Chainguard

Athena: 「パッチ・修正版」をクローズドコミュニティで囲い込む動きも





出典：https://www.chainguard.dev/athena

米国政府もフロンティアAIを活用した独自の脆弱性調整を宣言



出典：https://www.whitehouse.gov/releases/2026/07/white-house-launches-gold-eagle-initiative-for-unprecedented-cybersecurity-vulnerability-coordination/

CONCLUSION

顕在化するリスクは未知の脆弱性だけではない。
既知の脆弱性へのパッチ適用にも、
より一層の速度が求められる。

CHAPTER 04

企業のセキュリティ対策

with フロンティア AI の時代に必要な舵取り

現場で実践したい対策：コードと（安全な）パッチ

A 自社実装を診断・ペネトレで検査

高頻度かつ高精度で実施されることが求められる。現実的な費用で実施するために**AI活用も必須**と言える。モデルの性能・性質に応じたハーネスがあるとなお良し。



B パッチを迅速に取り込む

N-hour時代、パッチ速度そのものが防御力。パッチを当てるべき依存先の把握や最小化も有効。しかし、昨今はソフトウェアサプライチェーンリスクが飛躍的に顕在化し**「安全に」**にも気を払わねばならず…（後述）



B' パッチを「安全に」取り込む



0-day / N-day 脆弱性がなくとも、侵害が発生し得るリスク

2025 年以降、正規パッケージや配布経路そのものを侵害する**ソフトウェアサプライチェーン攻撃**が激化。広く依存されるライブラリが侵されれば、自社に脆弱性がなくても被害が連鎖する。

axios（2026/03侵害）
週間ダウンロード数
1億超 / 週

出典：https://blog.flatt.tech/entry/axios_compromise

A Takumi 診断・AI ペネトレーションテスト（7万円/月〜）



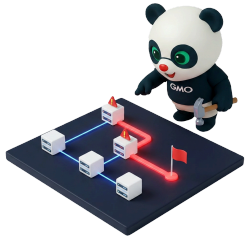
ホワイト/ブラックボックス診断

ソースコードに対する静的解析や、デモ環境の URL 一つで開始できる動的検証により、さまざまな脆弱性を検出。



脆弱性の自動修正

診断機能で検出した脆弱性の修正パッチをAIが自動作成し、人間は承認するだけ。テストコードも合わせて生成し、将来のデグレも防止。



AI ペネトレーションテスト

フロンティアAIが高度なテストを自律的に実行。複数の脆弱性を組み合わせた実際の攻撃シナリオで、検出に留まらず悪用可能性まで検証。

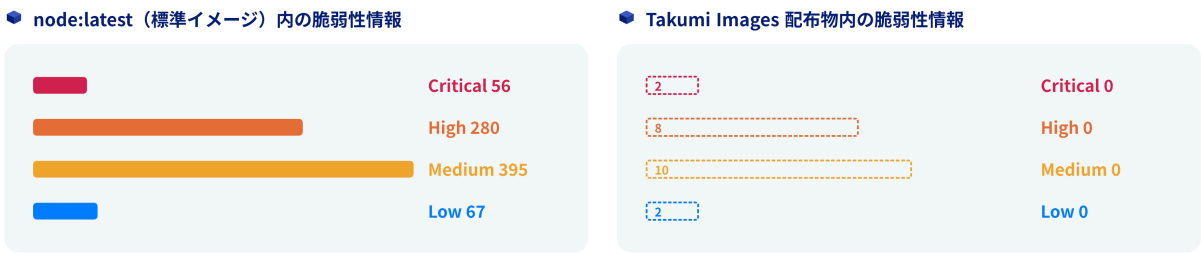
B Takumi Images（無料）



既知脆弱性を最小化し、マルウェア含有検査済みの コンテナイメージ

https://flatt.tech/takumi/features/images

脆弱性を最小化 | マルウェア検査済み | Dockerfile 1 行で導入・無料



※ 2026年7月6日時点。脆弱性検査は「Grype」による。点線内はスキャナが検知しうるが、悪用不可能と専門家が判定した脆弱性の数量

Takumi Guard（無料・管理機能有料）

悪意あるパッケージを検知・遮断し、サプライチェーン攻撃を防ぐ

https://flatt.tech/takumi/features/guard



12万 バージョン/日を検査

4,000万 DL/日をお守り

01

脅威は Mythos だけではないが、守りに転用可能

フロンティアAI全体で脆弱性の発見・悪用が実用段階に。他社モデル × **$25** で WordPress に認証不要 RCE すら現実となった。

02

パッチ適用の時間軸が変わった

AI がパッチ差分から数時間でエクスプロイトを自律生成し、**N-day は N-hour へ**。パッチ適用に関するこれまで以上の速度・工夫が求められる。

03

打ち手は「AI診断・ペネトレ」と「迅速・安全なパッチ」

AIによる高頻度・高精度・安価な診断・ペネトレと迅速で安全なパッチ運用を回す。**Flatt は Takumi 群**（診断・Images・Guard）でこの分野を最大限に支援。


GMO Flatt Security
Our Mission
エンジニアの背中を預かる
より多くのエンジニアがものづくりに集中できる社会を、
セキュリティ面からつくる会社