SECURITY REPORT · 2026-06-27

Claude Mythos 以後のセキュリティと、当方システムが取った対策

Mythos 級モデル一般提供時代における 8 層防御と Passive Safety の実装記録

Fable 5 / Mythos 5 | 8 層防御モデル | Passive Safety | 政府介入での緊急停止

竹下 剛

X：@fugusakate

想定発表時間 10 分

PROFILE

発表者

竹下 剛 たけした つよし

東証プライム上場企業（SIer / ベンダー）｜ PM / QA



本業	品質管理システムをフロントエンド〜インフラまでソロで開発・改善。 社内 40+ プロジェクトで利用されている。
個人 （副業承認済）	iOS / Android / Web アプリを個人開発。大学・中小企業と組み、 社会的意義のあるサービスをボランティアで開発中。
発信	X：@fugusakate（「生成AI × アプリ開発」を発信）

@fugusakate

AGENDA

本日のアジェンダ

01 **Mythos とは何か**
Project Glasswing と Mythos 級モデル

02 **特に X で話題「Fable 5 すごい」**
公開直後の反響とベンチマーク

03 **セキュリティの「勝ち筋」は変わった**
Passive Safety と AI エージェント

04 **当方システムで実装したこと（8 層モデル）**
8 層防御へのマッピング

05 **注目実装：自動停止トリガー 3 段階**
異常データアクセスへの自動応答

06 **フェーズ別ロードマップ**
止血 → 運用化 → 強化 → 定着

07 **最新動向 — 政府介入での「緊急停止」**
Fable 5 / Mythos 5 の停止措置（米国人以外）

08 **まとめ**
Mythos 以後の設計原則

POSITIONING

本発表の焦点 — 金融庁・日銀「フロンティア AI 要請」9 項目との対応

金融庁・日本銀行は 2026-05-22、金融機関等に短期的対応 9 項目を要請。本発表は、そのうち ⑥⑦⑧（リスクベース運用・パッチ以外の多層防御・停止への備え）に焦点を当てる。

#	金融庁・日銀 要請 9 項目（短期的対応）	本発表での該当箇所	焦点
1	経営課題として扱う（経営層の関与）	§3 勝ち筋 / §6 ロードマップ / §8 まとめ	○
2	優先サービス／IT システムを特定	§4 システム前提・8 層マッピング	○
3	特定資産の技術的負債を解消	§6 Phase A「止血」	○
4	パッチ適用の人的リソースを追加	（本発表では深掘りせず）	△
5	ベンダー維持保守契約を確認	（本発表の対象外）	−
6	**パッチ適用をリスクベースで運用**	§1 読み筋 / §3 勝ち筋①②	◎
7	**パッチ以外の対策も強化（多層防御）**	§3 Passive Safety / §4 8 層 / §5	◎
8	**優先サービス／IT の停止に備える**	§5 自動停止「遮断」/ §7 可用性	◎
9	外部との連携を維持・強化	§7 単一依存・外部要因 / 監視層	○

→ 本発表の主眼 ＝ ⑥ リスクベース運用・⑦ パッチ以外の多層防御（Passive Safety / 8 層）・⑧ 停止への備え。「侵入を前提に被害を最小化する」側に焦点。

出典：金融庁・日本銀行「フロンティアAIによる脅威変化を踏まえた金融機関等の短期的な対応」に係る要請（2026-05-22）

EXECUTIVE SUMMARY

サマリー — 3 行で言うと

01

Mythos は実務に入った

2026-05-29 に Opus 4.8、06-09 に Fable 5（Mythos 級の一般提供版）が公開。脆弱性の発見・解析が一気に民主化された。

02

勝ち筋は「止める・戻す」

「見つける」より、侵害されても被害が広がらない Passive Safety と、AI エージェントの権限管理が要になった。

03

そして政府が緊急停止した

06-12、米政府が輸出管理指令で Fable 5 / Mythos 5 を外国籍向けに停止。事実上、全世界で一時利用不能に（06-26 現在も継続）。

01 Mythos とは何か — Project Glasswing

防御目的に限定された最上位クラス

- 2026 年 4 月、Anthropic は「AI 時代の重要ソフトウェアを守る」**Project Glasswing** を開始（現在は約 150 社・15 か国超に拡大）。
- 中核モデル Mythos Preview は、脆弱性検出やペネトレーションテスト支援など**防御目的に限定**して提供。
- Opus 級の上に位置する「Mythos 級」という新しい能力ティア。サイバーや生物分野では桁違いの実力を持つ。



01 初期報告の数字と「能力の読み方」

約 50

参加パートナー

1 万件超

1 か月の高・重大度の脆弱性候補の報告

約 2 週間

平均パッチ期間（修正がボトルネック）

読み筋：「見つける」より「直す・止める・戻す」がボトルネック

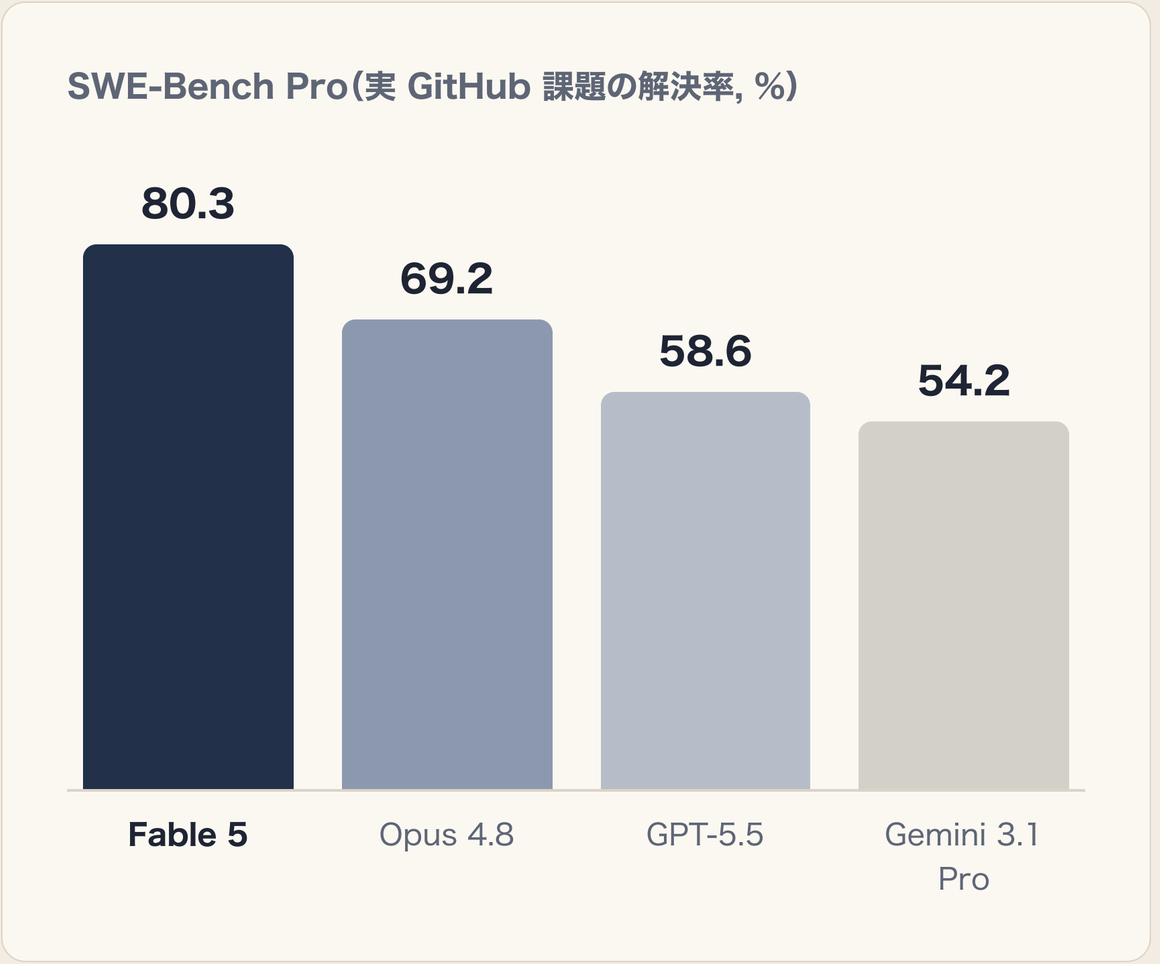
- 脆弱性候補は大量に出るが、パッチ適用が追いつかない。
- 英国 AI Security Institute も、実験環境と現実の防御環境には差があると明記。
- 脅威は現実的だが、過剰な恐怖と過小な備えのどちらも避けるのが正しい構え。

FABLE 5 IS HERE

02 Fable 5 — Mythos 級が一般提供に（6/9）

「うちが出せる中で最強」

- Mythos と同一ベースモデル。一般提供のため、サイバー・生物・化学などはガードレールで Opus 4.8 にフォールバック。
- ほぼ全ベンチで SOTA。SWE-bench Verified **95.5%**、長く複雑なタスクほど差が開く。
- Stripe：5,000 万行の Ruby 移行を **1 日**で完了（人手なら 2 か月超）。
- 価格は入力 $10 / 出力 $50（100 万トークン）。Mythos 5 はパートナー限定で分類器なし。



RECEPTION ON X

02 X での反響 —「すごい」を集めた実例

WordPress テーマを一発生成

Automattic の J. Marsland 氏が「スクショ + URL」だけで完全編集可能なブロックテーマを one-shot 生成 →「next level」。最も拡散した反応。

Minecraft クローンを約 20 分

「クローンを作って」の一言で、複数バイオーム・昼夜・鉱石・洞窟を生成。Windows 風 UI 再現デモも話題に。

コードの自己改善

S. Willison 氏：Datasette Agent の課題を解いた上に、依存ライブラリ側の 4 件の不具合まで発見・修正。

Pokémon FireRed をクリア

視覚のみの最小ハーネスで長期エージェント能力を実証。Slay the Spire は永続メモリで終盤到達 3 倍。

ただし冷静に：一部のバイラル・デモは「手作りでは」との指摘もある。派手な一発デモより、難しい実タスクでの再現性で評価したい。

THE WINNING MOVE CHANGED

03 勝ち筋①:Passive Safety という設計思想

> 攻撃が成立しない前提ではなく、**成立しても被害が拡大しにくい**ようにシステムを設計する。

01

本番デプロイは人手承認のみ

自動マージで本番に直接届かせない

02

大量データエクスポートで停止

閾値超でセッションを自動停止

03

バックアップ削除を検知・アラート

復旧経路を守る

04

CI/CD 改変の即時検知

パイプライン汚染を早期に発見

一般企業向けの Passive Safety は、高額な暗号技術より「運用ルール + 自動応答」から始められる。

THE WINNING MOVE CHANGED

03 勝ち筋②：AI エージェント＝新しい高権限利用者

Cursor・MCP・GitHub Copilot は「人間の補助ツール」ではない

実行権限を持つ「主体」として、人間の高権限ユーザーと同じ目線で扱う必要がある。

01 **人間の個人トークンを AI に渡さない**

AI 専用の最小権限クレデンシャルを発行する

02 **本番変更・データエクスポートは人間承認**

重要操作は必ず人のゲートを通す

03 **ツール呼び出しを監査ログに残す**

誰が・何を・いつ実行したかを追跡可能に

OUR SYSTEM

04 当方システムの前提 — なぜ他人事でないか

特定 IP 制限付きの Web アプリだが、構成要素は Mythos 以後の議論がそのまま当てはまる。

Bedrock LLM（5 経路）

プロンプトインジェクション対策の対象

GitHub Actions OIDC デプロイ

CI/CD 侵害・サプライチェーン

MCP / Cursor

AI エージェント権限の管理対象

Aurora / S3 の品質データ

大量エクスポート・データ窃取リスク

結論：規模は小さくても、Mythos 以後に問われる論点（LLM 入力・CI/CD・AI 権限・データ保護）を一通り抱えている。
だからこそ先回りして設計・検証した。

OUR SYSTEM

04 8 層防御モデルへのマッピング

1 アクセス境界
特定 IP 制限 Web アプリ

2 ID・認証
GitHub Actions OIDC / トークン分離

3 アプリ入力
プロンプトインジェクション対策

4 AI エージェント権限
MCP / Cursor を監査・最小権限化

5 CI/CD・供給網
パイプライン改変の即時検知

6 データ保護
大量エクスポート制御（API 単位）

7 バックアップ・復旧
削除検知・静的 S3 versioning

8 監視・自動応答
Passive Safety / 説明責任

現時点で 8 層すべてを実装・本番稼働済み。

PASSIVE SAFETY

05 注目実装：自動停止トリガー 3 段階

社内 SaaS では、ネットワーク全体の egress 遮断より**「API 単位のデータ持ち出し制御」**が現実的。

段階 1 ｜ 検知

異常な大量データアクセス／エクスポートを検知し、即アラート

ingest.job download（CSV）/ qep.download（Presigned URL）を監視

→

段階 2 ｜ 抑制

該当 API をレート制限・一時ブロックし、被害速度を落とす

正常業務は継続、持ち出しだけを絞る

→

段階 3 ｜ 遮断

閾値超でセッションを自動停止。本番変更は人手承認へ退避

侵害時に安全側へ倒す（fail-safe）

ROADMAP

06 3 日間ロードマップ

初日 — A

2 日目 — B

3 日目 — C

以降 — D

PHASE A

止血

緊急対応。最優先の穴を塞ぐ

PHASE B

運用化

対応を恒常的な運用へ落とし込む

PHASE C

強化

Security Hub / S3 versioning / Passive Safety を prod 投入

PHASE D

定着

新規大規模実装より、運用の習慣化に焦点

以後の焦点は、新規の大規模実装よりも「運用定着」。守りを習慣にすることがゴール。

BREAKING ・ 2026/06/26 時点

07 政府介入での「緊急停止」措置（米国人以外）

6/12（金）17:21 ET、米政府が輸出管理指令を発出。

Fable 5 / Mythos 5 を、米国内外を問わずすべての外国籍ユーザー（＝米国人以外）に対し停止。Anthropic 従業員の外国籍も含む。

発令元

米商務省（輸出管理）

根拠：国家安全保障

実質的な影響

全顧客で一時停止

選別が運用上困難なため

他モデル

Opus 4.8 等は影響なし

代替経路は機能

06-26 時点：復旧は未発表（停止 14 日目）。Anthropic 幹部は「近日中の再開に自信」とするも、公式の再開告知はなし。本節は 06-26 の公開情報に基づく追補。

WHAT HAPPENED

07 経緯と論点 — 提言が現実になった皮肉

タイムライン

- **06-04** ● Anthropic が「ブレーキペダル」提言（論考 When AI Builds Itself）
- **06-09** ● Fable 5 一般公開（Mythos 級の公開版）
- **06-12** ● 米商務省が輸出管理指令 → 全世界で停止（発端：脱獄の発見＋韓国通信の Mythos 利用・対中懸念）
- **06-18** ● 幹部「近日中に再開」と表明（ソウル会見）
- **06-25** ● 復旧の噂を公式否定（トラフィックは 0 と表明）
- **06-26〜** ○ なお停止中（14 日目）。当面の節目は 7/8（ID 検証）・8/1（EO 期限）

政府の言い分

ある脱獄手法（特定コードを読ませ欠陥を指摘させる類）が国家安全保障上のリスク、と判断。GPT-5.5 でも同種の能力は得られると Anthropic は指摘。

Anthropic の反論

「狭い潜在的脱獄を理由に、数億人が使う商用モデルを回収すべきではない。同基準を業界全体に適用すれば、新モデルの提供がすべて止まる」。規制は透明・公正・技術的事実に基づく法的手続きで、と主張。

WHAT IT MEANS FOR US

07 当方への含意 — 単一プロバイダ依存リスク

01

日本のユーザーは直撃対象

「米国人以外」への停止＝日本から使う我々はまさに対象。可用性は地政学とコンプライアンスに左右される。

02

可用性は一夜で消えうる

規制・係争・サイバー攻撃・輸出管理で最先端モデルは突然止まる。今回の停止も 2 週間超継続中。単一 API 依存は事業継続リスク。

03

マルチモデル / フォールバック設計

経路を複数持ち、停止時に自動で代替へ退避。Opus 4.8 等は無事、代替も台頭（例：Sakana『Fugu Ultra』）＝設計次第で被害を限定。

04

当方は継続稼働できる

当方システムは Bedrock 上の Sonnet / Haiku で動作。Mythos 級停止の直接影響を受けず、業務を止めない。

SUMMARY

まとめ — Mythos 以後の設計原則

01

一体で設計する

開発・運用・ID・バックアップ・AI エージェント・説明責任は分断せず一体で。単一ツールの導入では完結しない。

02

Passive Safety を前提に

「入られない」前提でなく「入られても広がらない」。検知 → 抑制 → 遮断の自動応答を運用に組み込む。

03

可用性は設計で守る

今回の緊急停止が示す通り、最先端モデルは一夜で止まる。マルチモデルとフォールバックで事業を止めない。

04

当方は先行実装済み

8 層防御と Passive Safety を社内 SaaS で先行実装・検証済み。Mythos 級停止下でも継続稼働できる。

SOURCES & CREDITS

詳細資料 と 出典

土台レポート

- Mythos セキュリティ統合レポート
- 8 層防御モデルと当方システム実装のマッピング
- 進捗管理：JIRA Epic（Phase A〜D）
- 2026-06-26 取得の公開情報で追補

主な出典（2026-06）

- 金融庁・日本銀行：フロンティア AI 要請 9 項目（2026-05-22）
- Anthropic 公式：Claude Fable 5 and Claude Mythos 5（06-09）
- Anthropic 公式：政府指令によるアクセス停止に関する声明（06-12）
- Anthropic Institute：When AI Builds Itself（06-04）
- CNBC / TechCrunch / VentureBeat / The Decoder（Fable 5 反響・性能）
- Korea JoongAng Daily・AWS 等（06-18〜26 の停止継続・復旧見通し報道）

※ Fable 5 / Mythos 5 のアクセス状況は流動的です。最新の公式情報をご確認ください。