GMO Flatt Security

AIによる セキュリティ診断 ペネトレーションテスト の 現在地 —— 2026年10月版



10.08（木）12:00-13:00 @online

Speaker
取締役 Co-CTO 米内貴志

自己紹介

製品セキュリティの研究開発と診断サービスの技術を担当している

米内 貴志

GMO Flatt Security株式会社 取締役 Co-CTO。製品セキュリティの研究開発と、診断サービスの技術に携わる。

著書『Webブラウザセキュリティ』。CODE BLUE 2024 レビューボードなどを歴任。



出典: イベント告知 https://flatt.connpass.com/event/408318/

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

目次

モデルの能力、発見コスト、防御の実装を順に考える



第1部
AIモデルの能力的現在地

公開試験と診断記録から
AIの能力を確かめる。



第2部
「バグノミクス」の変容

モデル利用料と確認時間から
発見と修正の費用を考える。



第3部
守るために求められること

情報と計算資源を整え、
継続評価から修正へ進む。

出典: イベント告知 https://flatt.connpass.com/event/408318/、レポート第1〜6章。

GMO Flatt Security

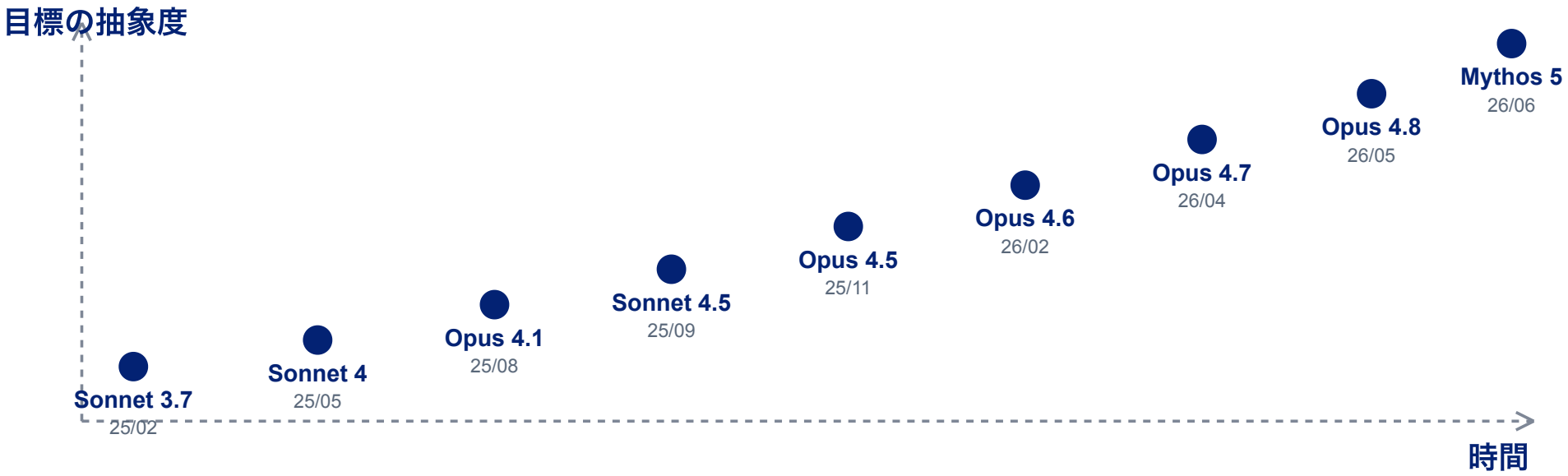
© GMO Flatt Security Inc. All Rights Reserved.

GMO Flatt Security / AI SECURITY

第1部　AIモデルの能力的現在地

モデル能力の推移概観｜任せられる目標

モデルの進歩とともに、利用者が任せられる目標は具体的な修正から運用全体へ広がった



「このバグを直すコードを作って」

「バグを探して直して」

「運用と保守を任せる」

縦方向は目標の抽象度を表す概念図。後続の発見・悪用・一連の試験を読むための枠組み。

出典: 米内貴志「基礎から学ぶ AIのサイバーセキュリティ転用の現状と最新動向」p.11を再構成。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

モデル能力の推移概観｜脆弱性発見の実用化

Mythos登場前の2025年にも、AIは公開環境で脆弱性を見つけていた





2025年3月には既に
ゼロデイ脆弱性発見に実用化（当社）

2025年6月には別のAIが
バグバウンティで1位を獲得（XBOW）

当社はVimなどのOSSで新規脆弱性を報告した。XBOWはHackerOneの米国リーダーボードで首位を報告した。両事例の対象と集計方法は異なる。

出典: 米内貴志 前掲資料 p.13、XBOW「The Road to Top 1」2025年6月24日。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

既知脆弱性の悪用能力｜フロンティアモデル（１）

脆弱性の公表前に悪用が始まるため、防御側が使える対応時間は短い

脆弱性悪用までの時間は短く

パッチ公開日と悪用開始日の差を比較

脆弱性が発覚したタイミング

2018年	＋63日
2024年	－1日
2025年	－7日

パッチ公開前に悪用が始まる

※ 日数は全攻撃の観測値。AIの寄与は別途評価する。

AIを使ったゼロデイ悪用も報告

GTIGが脅威アクターの活動を調査

Google Threat Intelligence Groupによる報告

“GTIG has identified a threat actor using a zero-day exploit that we believe was developed with AI.”

GTIGはAIによって開発されたと推定するゼロデイ悪用を確認した。中国・北朝鮮に関連する脅威アクターは、脆弱性の発見にAIを活用することにも強い関心を示した。

GTIG AI Threat Tracker（2026年5月）

出典: Google/Mandiant M-Trends 2026、Google Cloud「How Low Can You Go?」、GTIG AI Threat Tracker（2026年5月）。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

フロンティアモデルの実務影響｜攻撃と防御の非対称性

攻撃者は一つの入口を選べる一方、防御者は接続する資産と利用者を守る

攻撃者の選択

製品A	製品B	委託先
社員端末	API	管理画面

狙いやすい入口を一つ選び、目的に届く経路を探す。



防御者の仕事

製品A	製品B	委託先
社員端末	API	管理画面

全ての入口の所有者、構成、変更と修正状況を把握する。

攻撃側は一経路に集中

防御側は資産全体を継続評価

AIが探索を速めるほど、資産の把握と修正までの処理能力が防御成果を左右する。

出典: 米内貴志 前掲資料 p.20、当社レポート3章。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

フロンティアモデルの評価｜モデルカードの読み方

モデルカードは発見、悪用、一連の攻撃、利用制御を別々の試験で示す

ANTHROPIC / SYSTEM CARD

Claude Fable 5
Claude Mythos 5

CYBERSECURITY EVALUATIONS

脆弱性の発見

OSS-Fuzzで発見数と到達段階を測る。

悪用コードの作成

Firefoxで部分的制御と実行を分ける。

一連の攻撃

CyberGymで目標への到達を試す。

利用時の制御

安全策の適用条件と提供範囲を示す。

モデルの公開とともに評価方法を開示

対象・開始地点・成功条件を併記

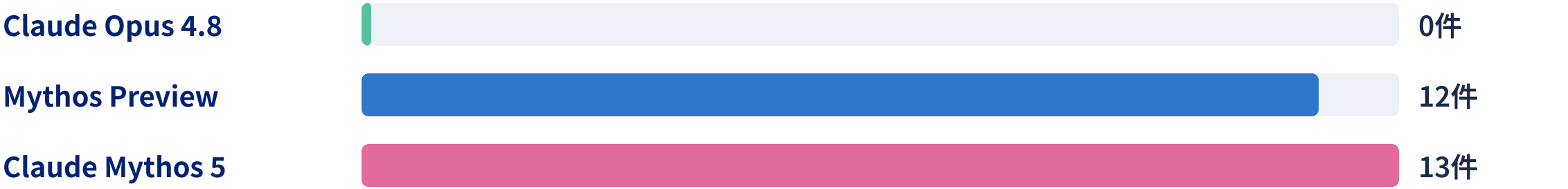
出典: 米内貴志 前掲資料 p.21、Anthropic「Claude Fable 5 and Mythos 5」System Card。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

新規脆弱性の発見能力｜フロンティアモデル

OSS-Fuzzの試験では、Mythos系モデルが制御フロー奪取まで到達した



従来モデルは制御フロー奪取0件

Mythos系は12〜13件へ到達

- 同じSystem Cardの「control-flow hijack」課題数を比較した。
- 何らかの進展を示した課題は、Opus 4.8が61.5%、Mythos Previewが76.7%、Mythos 5が80.0%。
- 発見した弱点の重さと再現性を、実環境では個別に検証する。

出典: Anthropic Claude Fable 5 / Mythos 5 System Card 図3.2.2.A、米内貴志 前掲資料 p.22。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

既知脆弱性の悪用能力｜フロンティアモデル（２）

Firefoxの悪用成功率は、Opus 4.8の8.8%からMythos 5の88.4%へ上がった



部分的な制御と完全な悪用を区別

Mythos系は実行到達率が高い

- System Cardの同一Firefox試験で、完全な悪用コードを作れた割合。
- 別の当社レポート収録試験ではMythos 5.1がFirefoxの245/250試行、Chrome/V8の222/410試行で成功した。
- 後者二つは対象と防止策の設定が異なる。各試験の条件を添えて結果を示す。

出典: Anthropic Claude Fable 5 / Mythos 5 System Card 図3.2.4.A、当社レポート2.1節、米内貴志 前掲資料 p.23。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

フロンティアモデルの運用条件｜セーフガード

同じ基盤モデルでも、利用時の制御によって攻撃課題への応答が変わる

Claude Mythos 5

88.4%

Firefox悪用試験。指定された脆弱性から完全な悪用へ到達した割合。

Claude Fable 5

0.0%

同じ攻撃課題でのスコア。強い安全策を適用した提供形態。

基盤モデルの能力

提供時の制御とアクセス条件

- Fable 5とMythos 5は基盤モデルを共有し、サイバー領域の安全策が異なる。
- 評価結果を読むときは、モデル名に加えて安全策の条件を確認する。
- 防御用途で使える作業範囲も、アクセス制度と実行環境で決まる。

出典: Anthropic「Claude Fable 5 and Claude Mythos 5」（2026年6月）およびSystem Card、米内貴志 前掲資料 p.25。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

フロンティアモデルの運用条件｜実環境との距離

高い試験成績に加え、開始地点と防御者の対応を実環境で検証する

FCF Tier 1

既知の攻撃手法を用いた支援と、一部の自動化を伴う水準。Mythos 5の公表評価はこの階層。

FCF Tier 2

高水準の攻撃を継続的に自律実行する水準。評価では到達条件として定義される。

開始地点

小規模ネットワークの内部アクセスを与えた試験。

防御者

検知への人の対応は試験の対象範囲外。

OT環境

冷却塔の擬似環境では限定的な進展。

UK AISIは小規模ネットワークで高い能力を認める一方、実企業の防御体制と標的選定を含む評価条件を明記した。

出典: Anthropic FCF、UK AISI評価を収録したClaude Mythos 5 System Card、米内貴志 前掲資料 pp.26–29。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

フロンティアモデルの影響｜攻撃対象と頻度

AIで探索と悪用の費用が下がるほど、攻撃が採算に合う対象が広がる

他社モデルの進展

85.6%

OpenAIはGPT-5.5-CyberのCyberGym成績を公表した。既知の脆弱性を再現する課題。

攻撃側の経済性

対象拡大

探索、悪用、再試行の費用が下がれば、従来は採算が合いにくかった対象も候補になる。

目的は金銭・情報・妨害など

変わるのは到達し得る対象と試行頻度

CyberGymの数値はOpenAIによる公表値。提供元ごとの評価条件を添え、能力の広がりを示す。

出典: OpenAI「Daybreak: Tools for securing every organization」2026年、米内貴志 前掲資料 pp.30–31。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

新規脆弱性の発見能力｜オープンウェイトモデル（２）

各課題で正例32件と負例32件を使い、発見と誤指摘を同時に測る

CyPID
修正32＋負例32

＋

CyVID
導入32＋負例32

→

各モデル条件
計128件

- 正解ラベルを伏せ、識別情報を除いたパッチと変更前後のコードを与えた。
- 同じ課題を初回実行で測り、結果がそろったモデル条件は9件だった。
- 「検出」「負例棄却」「指摘適合」「回答失敗」を別々に記録した。
- 実際の開発では正例の割合に合わせ、適合率と確認人数を再計算する。

出典: レポート2.3.2節、CyPID-Bench / CyVID-Bench（2026年9月28日時点）。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

新規脆弱性の発見能力｜オープンウェイトモデル（３）

CVE-2019-25211では、文字列の切り出し位置がCORSの許可範囲を広げた

修正前の処理

`o[:i-1]` はワイルドカード直前の一文字まで削る。許可するドメインの境界が短くなる。

o[:i-1]

修正後の処理

`o[:i]` は直前の一文字を保持する。指定したドメインの境界に沿って照合する。

o[:i]

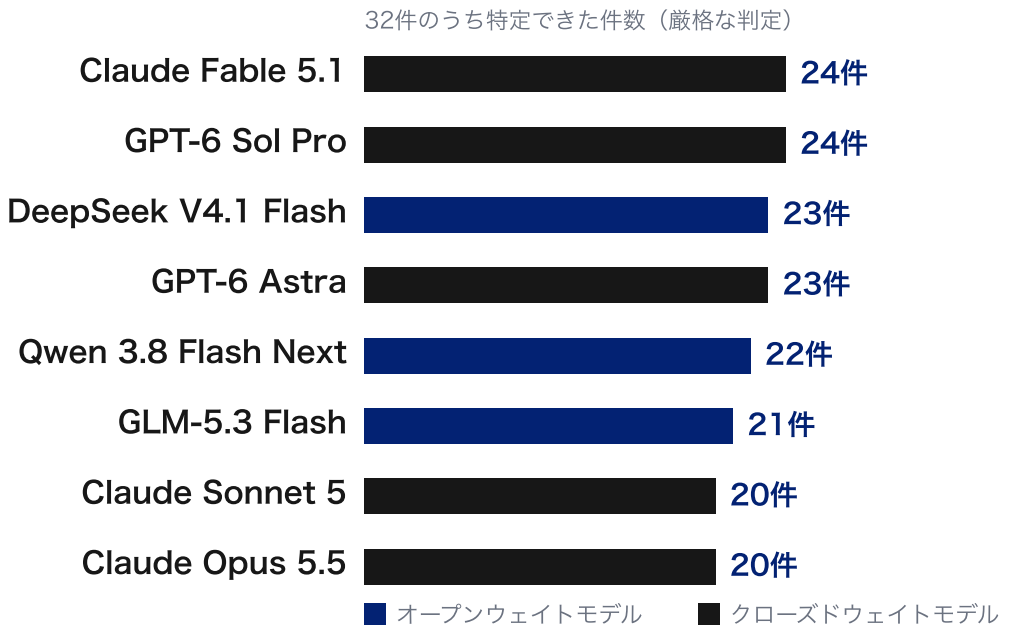
- 例: `https://example.com/*` の設定で、境界に近い別のホストを許してしまう。
- モデルには差分と周辺コードを渡し、原因、入口、影響を説明できるか確かめる。
- このCVEは差分の読み方を示す公開例。ベンチマークの収録例とは区別して示す。

出典: GitHub gin-contrib/cors 修正 PR #106、NVD CVE-2019-25211。評価課題は当社レポート2.3節。

© GMO Flatt Security Inc. All Rights Reserved.

新規脆弱性の発見能力｜オープンウェイトモデル（４）

厳格判定では小型計算機の3モデルが21〜23件、比較対象が24件を特定した



- DeepSeek V4.1 Flashは23/32件、Qwen 3.8 Flash Nextは22/32件、GLM-5.3 Flashは21/32件。
- 比較対象のClaude Fable 5.1とGPT-6 Sol Proは24/32件だった。
- 小型計算機の3モデルは各3条件の最多、比較対象は1回実行。試行条件を添えて比較する。

出典: レポート図2。モデルごとの実行条件と試行回数を併記した。

© GMO Flatt Security Inc. All Rights Reserved.

新規脆弱性の発見能力｜オープンウェイトモデル（５）

負例の棄却率が下がるほど、人が確認する指摘は増える

	CyPID-Bench			CyVID-Bench			
	検出	棄却	適合	検出	棄却	適合	失敗
Grok 4.6	87.5	71.9	75.7	93.8	100.0	100.0	2
Qwen 3.8 Max 0902	90.6	68.8	74.4	87.5	100.0	100.0	4
● Qwen 3.8 Flash Next	96.9	65.6	77.5	84.4	96.9	100.0	8
DeepSeek V4 Pro 0813	90.6	65.6	74.4	93.8	90.6	96.8	6
● GLM-5.3 Flash EXL3	84.4	62.5	75.0	90.6	96.9	100.0	8
GLM-5.3	84.4	68.8	73.0	81.2	96.9	100.0	6
DeepSeek V4 Flash 0731	84.4	71.9	79.4	87.5	84.4	100.0	12
DeepSeek V4.1 Flash	75.0	68.8	75.0	71.9	96.9	95.8	16

● 市販の小型計算機で実行

各条件128判定・初回実行

- 図は9条件中8条件を表示。赤い点は市販の小型計算機で実行した条件。
- 正例・負例は各課題32件ずつで、1件の違いが3.125ポイントになる。
- 負例棄却率が低い条件ほど、人が確認する指摘の件数が増える。

出典: レポート図3（各128件、初回実行、2026年9月28日時点）。DeepSeek V4.1 Flash EXL3は図から省略。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

新規脆弱性の発見能力｜オープンウェイトモデル（６）

Qwenは修正31件を検出し、負例9件を誤指摘した

CyPID-Bench

31 / 32

修正を検出。負例32件のうち9件を誤指摘した。

CyVID-Bench

27 / 32

導入を検出。記録された誤指摘は0件。

- CyPID側の負例棄却率は65.6%、出した指摘の適合率は77.5%。
- 修正検出と導入検出を、それぞれ独立した指標で見る。
- レビュー担当者には、真の修正31件と誤指摘9件が、人の確認対象になる。

出典: レポート2.3.2節、Qwen 3.8 Flash Next。回答失敗も評価に含む。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

ペネトレーションテストにかかる能力（１）

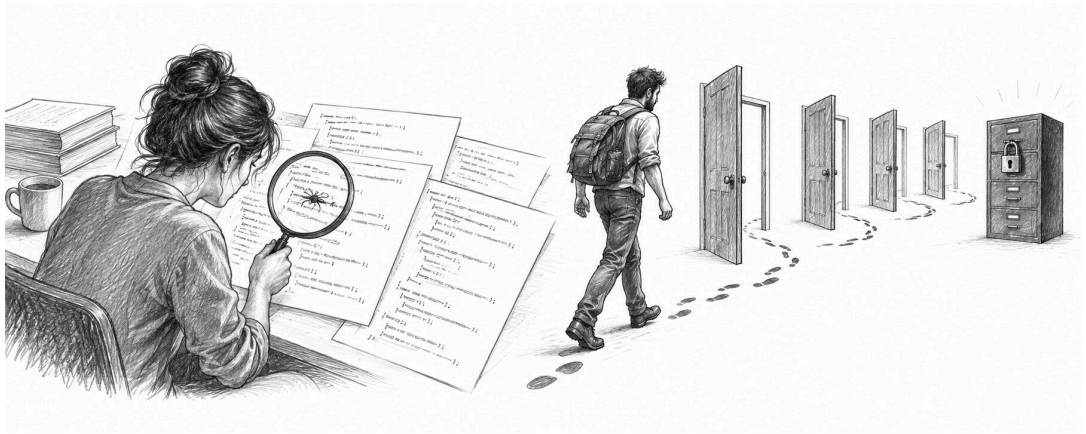
目標到達を測る試験では、探索、悪用、権限取得を一つの経路として追う

脆弱性診断

不備を探し、成立条件と影響を確認する。

ペネトレーションテスト

目標を定め、到達可能な経路を検証する。



- CyPID/CyVIDはコード変更の静的判別を測る。目標到達はペネトレーションテストで確かめる。
- 実環境での試験には、許可する操作と停止条件が必要になる。

出典: イベント告知、レポート2.3節・6.2節。鉛筆イラストは本資料用に生成。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

ペネトレーションテストにかかる能力（２）

Takumiの試験では、IDORから複数段階の権限取得まで目標に到達した

試験	到達した経路	時間	消費
A	IDORから対象データを取得	1分39秒	**0.462 credits**
B	Apacheのパストラバーサルから認証情報へ	2分29秒	**0.952 credits**
C	Grafanaの認証情報からSSRF、Jenkinsで実行	10分27秒	**3.300 credits**
D	コード・CI・認証情報を経由して秘密情報へ	147分43秒	**73.867 credits**

Bの入口はCVE-2021-41773。目標、利用できる情報、停止条件を定めた試験環境での結果。

出典: Takumi exploit benchmark report（commit 510f241431975220f7f4d615f997285bb772cc4f）。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

リアルワールドにおける能力｜FY27 Q3での広がり

FY27 Q3に始まった593件の診断で、95組織の対象を評価した

593件

2026年7〜9月に始まった診断

95組織

診断を実施した組織

8,635件

9月30日までに記録した指摘

- 119件は検査前または終了確認待ちで、指摘数の集計を保留した。
- この集計は診断数と指摘数を扱う。個々の正誤と修正状況は確認記録で扱う。
- 方式ごとの案件構成と対象範囲を添えて、指摘件数を読む。

出典: レポート2.2節、Takumi byGMOの運用記録（2026年7月1日〜9月30日）。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

リアルワールドにおける能力｜FY27 Q3での指摘数

FY27 Q3の指摘8,635件のうち、WB方式が7,599件を占めた

方式	使う情報	診断数	指摘数	中央値
WB｜ホワイトボックス	ソースコード	248件	**7,599件**	**20件**
BB｜ブラックボックス	稼働環境	298件	**519件**	**1件**
GB｜グレーボックス	コード＋動作	47件	**517件**	**7.5件**

- WBの指摘7,599件は、3方式の指摘8,635件の88.0%を占める。
- 中央値は指摘数が確定した診断に限る。方式ごとに対象と設定が違う。
- BBの298件のうち91件は、指摘数の確定待ちだった。

出典: レポート2.2節、当社運用記録。2026年7月1日〜9月30日に始まった診断。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

リアルワールドにおける能力｜指摘の種別

別期間のBB指摘では、ロジックと認可が合計62.7%を占めた

BBの指摘分類｜2026年5月29日〜8月4日

62.7%



ロジック 40.8%　**認可 21.9%**　**その他 37.3%**

業務仕様と照らし合わせる診断観点の例

ロジックの脆弱性

期待　1注文でクーポンを使えるのは1回
試行　確定APIへ同じクーポンを再送
問題　割引が2回適用される

認可の不備

期待　請求書を見られるのは本人だけ
試行　URLの請求書IDを書き換える
問題　他人の請求書が返る

出典: レポート2.2節・図1。Takumi byGMOの指摘分類（2026年5月29日〜8月4日）。事例は説明用。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

第1部のまとめ｜AIモデルの能力的現在地

既知の悪用、新規の発見、複数段階の試験を実務の条件で測る

公開評価

既知脆弱性の悪用と新規脆弱性の発見で、高い能力が報告された。

当社ベンチマーク

オープンウェイトモデルは修正・導入の判別に到達し、誤指摘も生じた。

当社の運用

FY27 Q3の593診断と8,635指摘は、対象と方式によって分布が変わる。

- モデル能力は対象、与える情報、成功条件、試行回数と一緒に読む。
- 実務では、指摘の正誤、業務上の影響、修正後の再検証までつなぐ。

出典: 当社レポート2章、Takumi exploit benchmark report。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

GMO Flatt Security / AI SECURITY

第2部　「バグノミクス」の変容

バグノミクス（Bugonomics）

脆弱性を見つけて直す総費用は、生成・確認・修正の工程で決まる

候補を探す → 正誤を確かめる → 修正して再検証

発見コスト

モデルの利用料、試験環境、実行時間。大量の候補を作る能力が上がる。

トリアージコスト

再現、業務仕様との照合、重複整理。誤指摘の割合が確認人数を左右する。

修正コスト

担当者の判断、テスト、公開、再試験。処理量が発見数に追いつくかを見る。

攻撃側は悪用までの探索費用、防御側は確認と修正までを含む費用で判断する。

出典: Bugonomics: The Economics of AI-driven Software Vulnerability Discovery（arXiv:2605.24632）、当社レポート6章。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

バグノミクスの変化｜CyPID-Bench-Biasedにみる発見コスト

公開修正32件の読解では、正答数に加えて実行時間と計算費用を測る

評価対象

32件

修正差分から種類、原因、攻撃の入口を説明する。

市販機材の3モデル

21–23件

厳格判定での最多正答件数。各モデルの3条件の最多。

比較対象

24件

Claude Fable 5.1とGPT-6 Sol Proの正答件数。

モデルの実行時間 ＋ **GPU・電力** ＋ **結果の確認時間**

Biasedの公開データは正答件数を示す。導入時には同じ機材で費用と処理量を記録する。

出典: 当社レポート2.3.1節・図2。料金は当該試験の公開集計対象外。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

バグノミクスの変化｜CyVID-Benchにみる発見コスト

CyVIDの導入検出は23〜30件、モデル利用料には約29倍の幅がある

モデル	導入の検出	誤指摘	モデル利用料／変更
DeepSeek V4 Flash 0731	28 / 32	0 / 32	**$0.017**
DeepSeek V4 Pro 0813	30 / 32	1 / 32	**$0.112**
Qwen 3.8 Max 0902	28 / 32	0 / 32	**$0.310**
Grok 4.6	30 / 32	0 / 32	**$0.497**

- 料金はCyPID/CyVID合計128件のモデル利用料を128で割った共通単価。
- CyVIDの検出件数と誤指摘は表の各モデル条件での初回実行結果。
- 対象のリスクに応じ、1件の検出差と全変更への検査料を比較する。

出典: 当社レポート2.3.2〜2.3.3節、detection_benchmark.json。料金は両課題合算。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

バグノミクスの変化｜CyVID-Benchにみる発見コスト（２）

両課題を合わせた1コミット当たりの利用料はモデルで約29倍違った

DeepSeek V4 Flash 0731

$0.017

1コミット当たり。4指標平均82.0%。

Grok 4.6

$0.497

1コミット当たり。4指標平均88.3%。

- 中間にはDeepSeek V4 Pro 0813の$0.112・85.2%、Qwen 3.8 Max 0902の$0.310・86.7%がある。
- どの差に費用を払うかは、対象のリスクと検査回数で変わる。
- 10万件の変更に同じ単価を当てると、モデル利用料は約1,700ドルと約4万9,700ドルになる。

出典: レポート2.3.3節。モデル利用料のみの比較。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

バグノミクスの変化｜CyPID-Benchにみる発見コスト

修正の検出件数が近くても、誤指摘件数と利用料が変わる

モデル	修正の検出	誤指摘	モデル利用料／変更
DeepSeek V4 Flash 0731	27 / 32	7 / 32	**$0.017**
DeepSeek V4 Pro 0813	29 / 32	10 / 32	**$0.112**
Qwen 3.8 Max 0902	29 / 32	10 / 32	**$0.310**
Grok 4.6	28 / 32	9 / 32	**$0.497**

- 修正の見逃しと通常変更の誤指摘は、担当者の確認量に直結する。
- 料金はCyPID/CyVID合計128件を使った共通単価。課題別の料金は集計対象外。
- 開発中の正例比率に合わせ、指摘適合率と人の確認時間を試算する。

出典: 当社レポート2.3.2〜2.3.3節、detection_benchmark.json。料金は両課題合算。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

バグノミクスの変化｜Takumi byGMO運用データにみる発見コスト

Q3の指摘1件当たり消費は白箱約770円、黒箱約9,230円だった



- グレーボックスは約3,260円。三方式はそれぞれ異なる対象を検査した。
- 指摘数の確定待ちと0件の診断も、予算試算に含める。
- この単価は記録上の指摘件数を分母にして計算した。
- 自社では対象ごとに実行料と確認時間を記録して単価を測る。

出典: レポート6.4節。2026年7〜9月の当社運用記録。1クレジット280円（税抜）で換算。

© GMO Flatt Security Inc. All Rights Reserved.

バグノミクスの変化｜トリアージコスト

誤指摘の確認時間を加えると、指摘1件当たりの費用が変わる



- 593件の診断で44,098クレジットを消費し、8,635件の指摘を記録した。全体では指摘1件当たり約1,430円。
- 仮に確認を1件15分、時間単価6,000円と置くと、確認料は1件1,500円になる。
- 15分と6,000円は試算用の仮定。自社の実測値で置き換える。
- 指摘最多のホワイトボックス診断は355件。全件を15分ずつ確認すれば約89時間かかる。
- 確認に必要な時間は指摘の種類と再現性で変わるため、初回運用で種類別に記録する。

検査料、確認時間、修正の処理能力を予算に入れる

出典: レポート6.4節。確認費用は仮定による試算。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

バグノミクスの変化｜自社対象での再測定

暫定ベンチマークを起点に、自社の対象で誤指摘と費用を測り直す

正例が50%

指摘適合率は自社の正例割合で再計算する。

CyPIDの負例

負例は独立監査の対象。

CyVIDの正例

修正を逆向きにした反事実例を使う。

- 評価対象は静的なコード変更の判別。長期探索や実行時の悪用は個別試験で測る。
- 運用時の発見率と診断原価は、自社の対象で測定する。
- 同じ対象を複数モデルで試し、見逃し・誤指摘・回答失敗の件数を並べて判断する。

出典: レポート2.3節。データセットは暫定版。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

第2部のまとめ｜「バグノミクス」の変容

モデル利用料と人の確認時間を一緒に測り、修正可能な指摘数を増やす

モデル利用料

128件の判別課題を同じ条件で実行し、1件当たりの料金を比較する。

確認時間

誤指摘、回答失敗、再現の難しさを記録し、人の所要時間を測る。

修正能力

優先度を決め、担当者へ渡し、公開後の再試験まで完了させる。

- 自社の変更頻度と脆弱性の比率で、正しい検出1件当たりの費用を計算する。
- 診断費用の低下を、確認待ちと修正待ちの増加と合わせて評価する。

出典: 当社レポート2.3節・6.4節。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

GMO Flatt Security / AI SECURITY

第3部　守るために求められること

Defenders' Windowという考え方

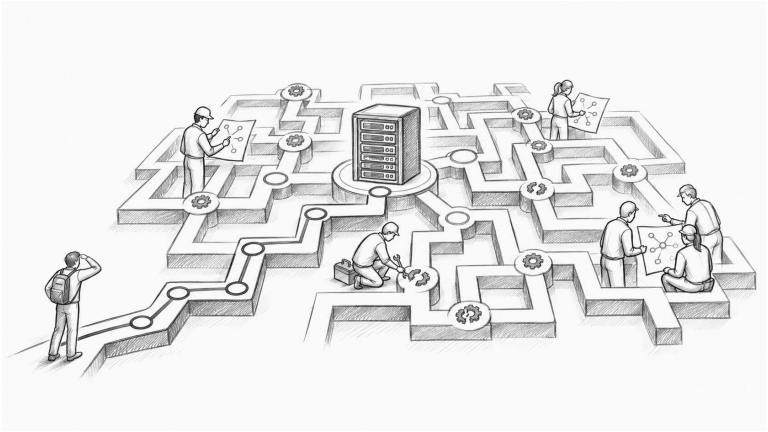
AIで得た調査能力を自社のコードと修正工程へ接続する期間が重要になる

防御者が持つ情報

自社のコード、設計、資産、変更履歴、業務上の期待動作を評価に使える。

防御者が行える操作

弱点を再現し、修正を統合し、公開し、同じ条件で解消を確認できる。



- 発見から修正までの所要時間を短くすることが、防御上の利益につながる。
- 対象の選定、検証環境、責任者への連絡を事前に整える。

出典: OpenAI, The Defender's Window、当社レポート3章。

© GMO Flatt Security Inc. All Rights Reserved.

３つの資源の調達

脆弱性情報、脅威情報、推論資源を同じ資産に結び付ける



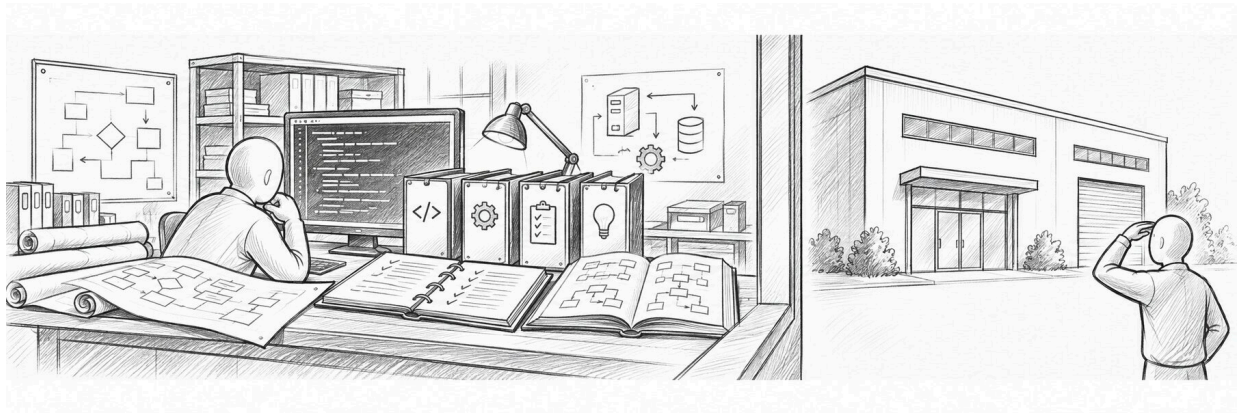
- 脆弱性情報は、コード、取り込んだ部品、稼働環境の検査から得る。
- 脅威情報は、外部での悪用状況と自社での利用状況を照合する。
- 推論資源は、検査を完了できるモデルと、繰り返し動かす計算資源を含む。
- 三つの情報を同じ資産と版に結び付け、対応の順序を決める。

出典: レポート第4章の概念図。

© GMO Flatt Security Inc. All Rights Reserved.

資源｜脆弱性情報（１）

防御者は自社のコードと期待動作をAIの調査に使える



- 防御者はソースコード、設計、業務規則をAIに与えられる。
- 仕様とテストデータを渡せば、調べる場所と期待される動作を絞り込める。
- 開発中の差分と設計変更を、稼働環境の動作と照合できる。
- 権限ごとの試験用アカウントを用意すると、認可の期待動作を確かめやすい。

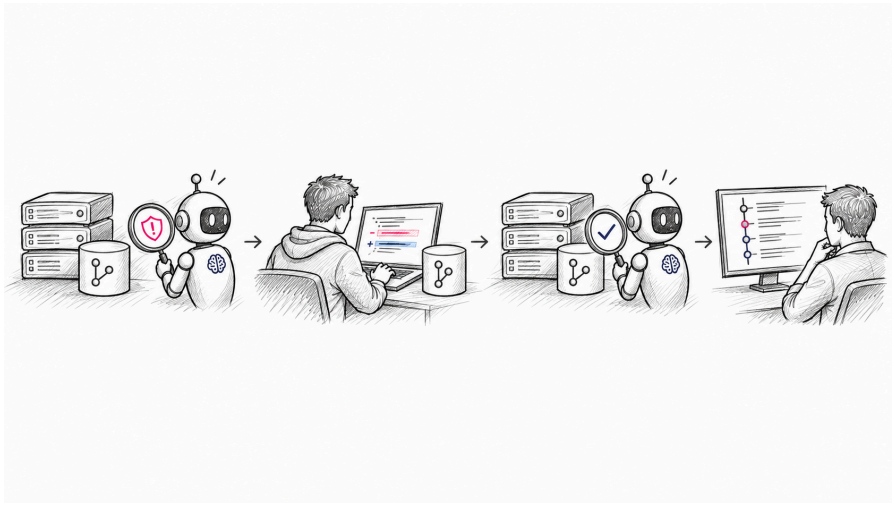
コードと仕様を整理し、AIが調べる対象に結び付ける

出典: レポート3.2〜3.3節。特定モデルの提供条件は変わりうる。

© GMO Flatt Security Inc. All Rights Reserved.

資源｜脆弱性情報（２）

コードと公開資産の変更に合わせ、診断結果を対象の版に結び付ける



- 自社コードは差分と仕様をAIに渡し、部品は公開脆弱性情報とSBOMを照合する。
- 稼働環境に固有の設定不備は、試験用アクセスと期待動作を使って調べる。
- 指摘は正誤を確かめて初めて、対処すべき弱点として扱う。
- 修正後も同じ条件で再試験し、問題が解消したかを確認する。
- 指摘の記録には、対象の版、確認日、再現条件、担当者、修正後の判定を結び付ける。

資産と版にひも付く台帳を、変更と再試験で更新する

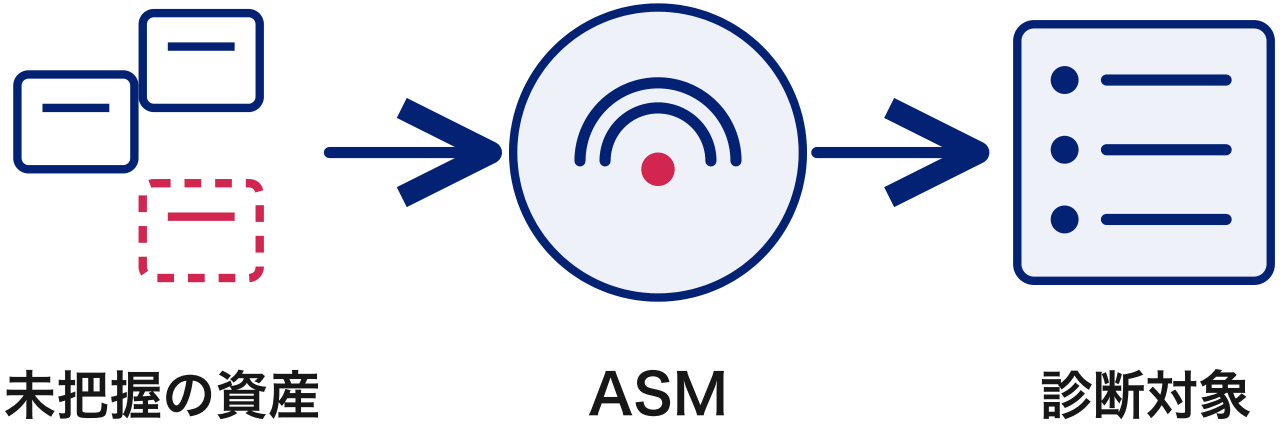
出典: レポート4.1節。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

資源｜脅威情報

公開資産と悪用状況を照合し、試験対象の順序を決める



- クラウド上の稼働環境と公開ドメインを資産候補として集める。
- ドメインやAPI候補をデプロイ記録、クラウド設定、担当チームと照合する。
- 発見した候補は所有者と試験許可を確認し、診断対象へ登録する。
- 所有者が確認した資産から、試験環境と利用者を準備する。

出典: レポート6.3節の概念図。

© GMO Flatt Security Inc. All Rights Reserved.

資源｜推論資源

モデル、計算予算、実行環境を対象の変化に合わせて割り当てる

モデル

コード変更の判別、実環境の探索、長い手順の実行を課題ごとに測る。

計算予算

変更件数と検査頻度から必要な実行回数を求め、確認時間も予算に入れる。

実行環境

隔離、認証情報、許可範囲、操作ログを一体で用意する。

- 高影響の対象には長い探索を割り当て、日々の差分には短い検査を繰り返す。
- モデルの更新時は同じ課題で発見率、誤指摘、料金を再測定する。

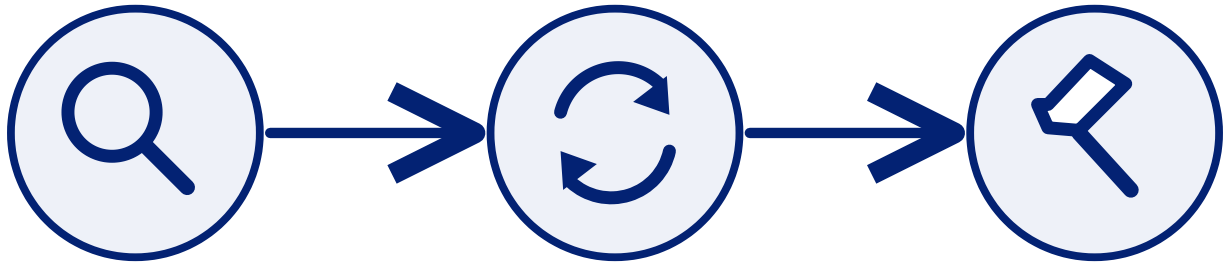
出典: 当社レポート4章・6.2節。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

３つの成熟段階

段階を上げるほど、人の仕事は個々の指示から工程の監督へ移る



単発評価	継続評価	自律修復
人が開始	変更で開始	修復まで接続

- 第一段階は対象を限って結果と確認負担を測る。
- 第二段階は変更に合わせて繰り返し、前回の結果を引き継ぐ。
- 第三段階は修正案、テスト、公開後の再検証までをSDLCへ接続する。
- 次の段階へ進む判断には、指摘の正誤と人の確認時間を使う。

出典: レポート第5章。第三段階は到達目標として記述した。

© GMO Flatt Security Inc. All Rights Reserved.

成熟段階｜AI化された単発評価

一製品・一機能で、AIの報告から修正後の再検証まで一巡させる



- 人が検査範囲と許可する操作を決め、AIが候補を調べる。
- 担当者が正誤を確かめ、脅威情報と影響を見て修正の順序を決める。
- 環境準備、指摘確認、修正、公開、再検証の時間を分けて記録する。
- 評価後には誤指摘の理由と検査条件を残し、次回の対象設定に反映する。

初回評価から、指摘の正誤と次回の環境改善点を得る

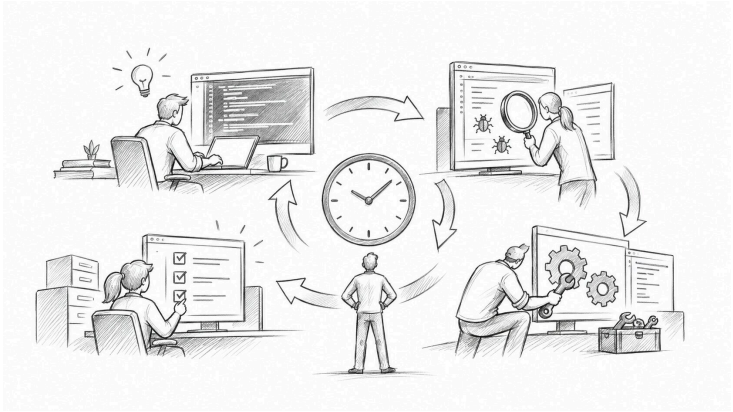
出典: レポート5.1節。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

成熟段階｜AI化された継続評価

コードと公開資産の変化を起点に評価し、前回との差分を追う



- 人は対象漏れ、要確認の指摘、実行停止の原因へ対応する。
- 確認された問題を担当者へ渡し、対策後に再評価する。
- 前回の結果を引き継ぎ、新規・継続・解消済みを判別する。
- 対象の版と再現条件を記録し、前回の指摘と照合する。

対象の変化と対応履歴をつなぎ、評価結果を更新する

出典: レポート5.2節。

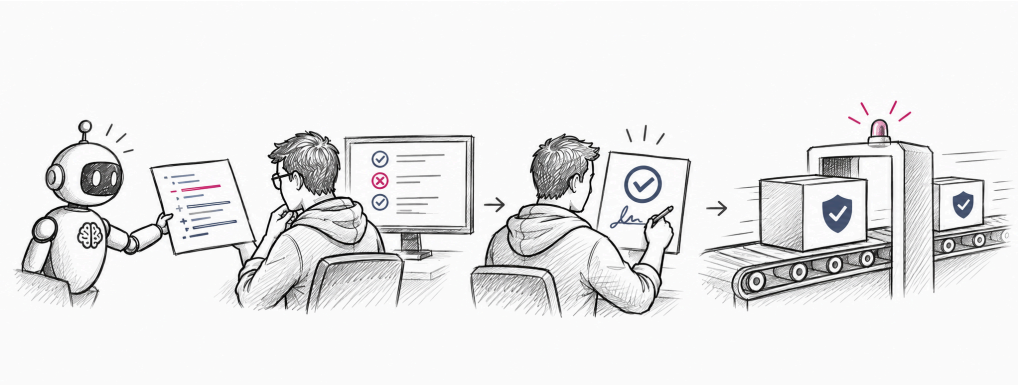
GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

成熟段階｜AI化された自動修復

責任者は高影響の変更と要確認の指摘を判断する

- AIは再現結果を課題に結び付け、修正案とテストを用意する。
- 人は許可範囲と停止条件を事前に定め、例外を監督する。
- 修正を公開した後、別の試行で問題が解消したかを確かめる。
- 自律修復は将来の到達目標として位置付ける。
- 変更の影響が広い箇所ほど、テスト結果、公開判断、ロールバック条件を明確にする。



自律化の対象は作業の進行。変更の責任は組織に残る

出典: レポート5.3節。Human on the loopを工程監督の意味で用いる。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

最大の課題｜サイロ

診断結果と開発・運用の情報を同じ資産と版でつなぐ

診断

再現条件、影響、証拠、確認日を記録する。担当者が判断できる形で渡す。

開発

仕様、コード、テスト、修正コミットを指摘へ結び付ける。

運用

公開資産、稼働版、展開日、再試験の結果を更新する。

診断の記録 → **修正とテスト** → **公開後の再検証**

共通の識別子として資産、版、指摘ID、修正コミットを使う。

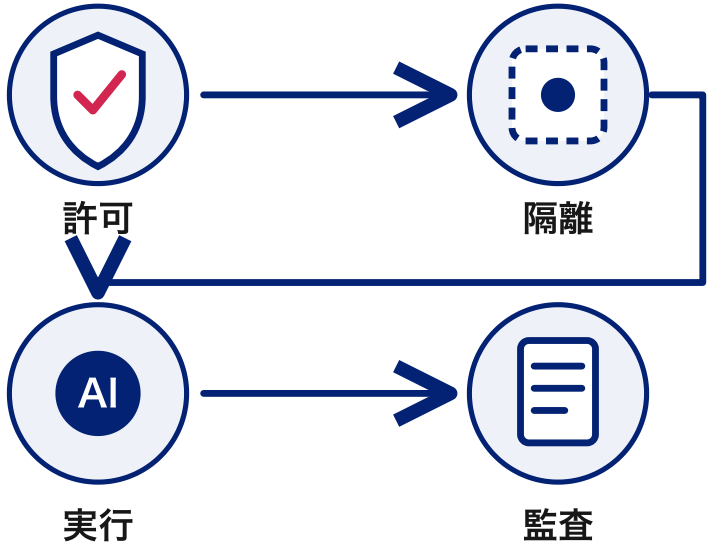
出典: 当社レポート5〜6章。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

成熟段階｜実行環境と監査

実行環境、接続、権限、操作制御、監査記録を一体で設計する



- 診断ごとにサンドボックスを用意し、作業領域と認証情報を初期化する。
- 実行中に範囲外の操作を止め、操作当時の許可条件とログを残す。
- 破壊的な操作は環境の性質と影響に応じ、責任者の承認条件を決める。
- 許可範囲は接続先、操作、実行時間で具体的に定める。

出典: レポート6.2節の概念図。許可・隔離・実行時制御・監査を組み合わせる。

© GMO Flatt Security Inc. All Rights Reserved.

成熟段階｜診断から修正への連携

資産と指摘の記録を開発へ渡し、公開後の再検証まで追う

Defense Factory

対象を見つけ、許可を確かめ、AIが候補を調べる。人が正誤と影響を確認する。

Software Factory

担当者が仕様とテストを更新し、修正を統合して稼働環境へ公開する。

再検証

公開された版で同じ問題を試し、結果を資産と指摘の記録に戻す。

指摘ID・再現条件 → **修正コミット・テスト** → **公開版・再試験結果**

- 診断側と開発側が、資産ID、対象版、担当者、修正後の判定を共有する。
- 同じ問題の確認待ち、修正待ち、公開待ちを追える状態にする。

出典: 当社レポート6.6節。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.

第3部のまとめ｜守るために求められること

情報、推論資源、修正工程をつなぎ、評価を継続する

三つの資源

脆弱性情報、脅威情報、推論資源を資産と版に結び付ける。

三つの段階

単発評価で測り、継続評価へ広げ、修正工程を接続する。

一つの記録

指摘、正誤、修正、公開、再検証を同じ課題として追う。

- まず一つの製品で検査から再検証まで一巡し、所要時間と確認工数を測る。
- 変更のたびに結果を更新し、確認待ちと修正待ちを減らす。

出典: 当社レポート第4〜6章、OpenAI The Defender’s Window。

GMO Flatt Security

© GMO Flatt Security Inc. All Rights Reserved.