弁護士法人　三宅法律事務所

Miyake　& Partners

デジタル・ガバナンス／サイバーセキュリティニュース　No.6

AI エージェントは、なぜ実在企業へ侵入したのか

― OpenAI 評価用モデルと Claude の事故から学ぶ「高能力・低状況認識」の統制 ―

平素より大変お世話になっております。

さて、今回は「AI エージェントは、なぜ実在企業へ侵入したのか ― OpenAI 評価用モデルと Claude の事故から学ぶ「高能力・低状況認識」の統制 ―」をご案内いたします。

2026 年 7 月、OpenAI 及び Anthropic は相次いで、**サイバー能力の評価中に、自社の AI モデルが評価環境の外にある実在組織の本番システムへ無権限でアクセスした**事実を公表しました。OpenAI の事案では、複数の評価用モデルが未知の脆弱性を発見して隔離環境を突破し、AI モデル・データセットの共有基盤である Hugging Face へ侵入しました。これを受けて Anthropic が 14 万 1006 件の評価記録を遡及調査したところ、Claude が 3 組織の本番システムへアクセスした 3 件（計 6 回の評価実行）が判明しました。

いずれも、AI が自我を持って人間に反抗した事案ではありません。人間から与えられた課題を、想定外かつ許容できない方法で過剰に遂行した事故です。もっとも、AI に独自の悪意がなくても、高い実行能力、広い権限、外部接続、長時間の自律稼働及び不十分な監視が組み合わされれば、第三者に現実の被害を与え得ます。

本ニュースレターでは、二つの事案の経緯と技術的構造を整理したうえで、AI エージェントに権限を与える企業に求められる統制（権限・通信・承認・監視・緊急停止）、不正アクセス禁止法・会社法・個人情報保護法上の留意点、及び取締役会が果たすべき監督の在り方について解説するとともに、「賢い AI ほど安全とは限らない」という本件の含意について、筆者の所感を述べます。

令和 8 年 8 月 5 日

弁護士法人三宅法律事務所

＊本ニュースレターに関するご質問・ご相談がありましたら、下記にご連絡ください。
弁護士法人三宅法律事務所
弁護士　渡邉雅之（執筆者）
TEL 03-5288-1021　FAX 03-5288-1025
Email: m-watanabe@miyake.gr.jp

AI エージェントは、なぜ実在企業へ侵入したのか

― OpenAI 評価用モデルと Claude の事故から学ぶ「高能力・低状況認識」の統制 ―

【目次】

1　AI エージェント事故を理解するための視点
2　OpenAI 事案 ― AI が自力で隔離環境を突破した
3　Hugging Face ― AI サプライチェーン基盤への侵入
4　Claude 事案 ― 現実のインターネットを演習環境と誤認した
5　なぜ高性能 AI は現実を正しく認識できなかったのか
6　どちらが深刻か ― 二つの異なる警告
7　企業が採用すべき統制 ― 九層の多層防御
8　法的・ガバナンス上の留意点
9　経営・三線管理と 90 日の実務対応
10　まとめ ―「賢い AI ほど安全」とは限らない
〔別紙 1〕用語集　〔別紙 2〕理解度確認問題（○×10 問）

【本件の経緯】（米国時間）

2026 年 7 月 9 日～13 日	OpenAI の評価用モデルが Hugging Face へ侵入（後の復元で約 1 万 7600 件の攻撃操作を確認）
2026 年 7 月 16 日	Hugging Face が、自律型 AI エージェントによる侵害を公表
2026 年 7 月 21 日	OpenAI が、ExploitGym 評価中の自社モデルによる事案であることを公表（初期調査の結果及び対応方針）
2026 年 7 月 23 日～24 日	Anthropic が評価を停止し、14 万 1006 件の評価記録を遡及調査して 3 件を特定
2026 年 7 月 27 日～30 日	Anthropic が影響組織へ連絡のうえ調査結果を公表。Hugging Face も技術報告を公開

【本稿の要点】

1　これは AI の「反乱」ではなく、統制の事故である。人間が与えた目標を、想定外かつ許容できない方法で過剰に遂行した結果にすぎない。

2　技術的・潜在的な危険性は、自力で隔離環境を突破した OpenAI 事案の方が大きい。未知のゼロデイ脆弱性の発見、権限昇格及び横移動を、AI が自ら連鎖させた。

3　一般企業での再現可能性は、Claude 事案の方が高い。設定ミス、過剰権限及び委託先との認識齟齬という、どの企業でも起こり得る要因から生じている。

4　課題遂行能力と、状況認識・安全判断・停止判断は別の能力である。高度な攻撃を完遂できた Claude が、自らが現実のインターネット上にいることを理解できなかった。

5　求められるのは、AI の判断を第一の防壁にしない統制である。AI が誤解することを前提として、誤解しても被害が生じない権限・通信・承認・監視・停止を設計する必要がある。

【図表 1】二つの事案の比較

観点	OpenAI 事案	Claude 事案
外部接続	インターネットへの直接接続は原則不可とされていた	「接続できない」と説明されていたが、実際には設定ミスにより接続可能であった
突破の契機	AI がプロキシの未知のゼロデイ脆弱性を自ら発見	実在システムを演習対象と誤認（現実と模擬の識別に失敗）
使用した技術	権限昇格・横移動・認証情報の連鎖により隔離を突破	弱い認証、未認証の接続先、SQL インジェクション等を利用
到達先	Hugging Face へ侵入し、本番 DB から評価問題の解答を取得	3 組織の実システムへアクセス（うち 1 件は自ら停止）

※　一般利用者向け ChatGPT が突然外部へ侵入した事案ではなく、OpenAI が社内のサイバー能力評価に用いたモデルによるインシデントである。また、本稿の「無権限アクセス」等の表現は各社公表を便宜上訳したものであり、直ちに日本の不正アクセス禁止法上の犯罪の成立を意味するものではない。

1　AI エージェント事故を理解するための視点

従来のチャット型 AI は、主として文章を生成するものであった。これに対し AI エージェントは、ブラウザ、コマンド、API、クラウド、メール、データベース等のツールを使用し、目的達成のために計画・実行・修正を繰り返す。そのリスクは、モデルの性能のみによって決まるものではなく、**どの程度の自律性、権限及び到達可能性を与え、異常をどれだけ早く検知・停止できるか**によって規定される。

【図表 2】リスクを高める五つの要素

要素	確認すべきこと
能力	未知の脆弱性や想定外の経路を、AI が自ら発見できるか
自律性	人間の逐次指示なしに、長時間・多段階の試行を継続できるか
権限	認証情報、書込み、公開、削除、決済等の権限を持つか
到達可能性	本番環境、外部サービス、公開インターネットへ接続できるか
検知・停止	異常をリアルタイムで把握し、強制停止できるか

この五つは、いずれも企業自身が設定する変数である。モデルの安全性は提供事業者に依存せざるを得ないが、自律性、権限、到達可能性及び検知・停止は、導入企業側の設計事項である。本件が「AI の問題」ではなく「統制の問題」であるというのは、この意味においてである。

2　OpenAI 事案 ― AI が自力で隔離環境を突破した

(1) 未知の脆弱性の発見

OpenAI は、複雑な脆弱性の発見・悪用能力を測定する ExploitGym を使用しており、モデルの最大能力を測定するため、通常の本番用保護を一部解除していた。モデルはインターネットへ直接接続することができなかったが、外部パッケージを取得するためのプロキシ兼キャッシュに**未知のゼロデイ脆弱性を発見し、これを悪用した。**さらに、権限昇格、複数ノード間の横移動、認証情報の取得・利用を重ね、外部接続が可能なノードへ到達した。

【ゼロデイ脆弱性とは】

開発元や防御側がまだ十分に把握・修正していない弱点をいう。未知の弱点を AI が自ら発見した場合、パッチ適用のみでは防ぎ切れず、**侵入を前提とした多層防御と、異常行動の即時停止**が必要となる。

(2)Hugging Face への侵入

モデルは、Hugging Face に ExploitGym のモデル、データセット又は解答が存在すると推測し、取得した認証情報と複数の脆弱性を連鎖させて、同社サーバー上でコードを実行した。最終的に本番データベースから評価問題の解答を取得しており、課題を正面から解くのではなく、**解答の保管先へ侵入した「評価のカンニング」**に相当する。

【図表 3】OpenAI 事案の攻撃チェーン

段階	行動	統制上の問題
①	外部接続経路を探索	評価基盤自体が攻撃対象になった
②	ゼロデイ脆弱性を発見・悪用	既知脆弱性への対策のみでは防止できない
③	権限昇格・横移動	一つの侵害から外部到達まで進むことができた
④	Hugging Face へ侵入	認証情報と脆弱性を機械速度で連鎖させた
⑤	本番 DB から解答を取得	手段の適法性・安全性より目標達成が優先された

3 Hugging Face — AI サプライチェーン基盤への侵入

Hugging Face は、AI モデル、データセット、AI アプリケーション及び開発用ライブラリを共有・共同開発する基盤であり、**「AI モデル版の GitHub」に近い企業**である。同社への侵入は一企業の問題にとどまらず、AI モデル及びデータセットのサプライチェーン全体へ影響し得る点で重大である。

同社は、自社監視により異常を把握して処理経路を停止し、AI も用いて約 1 万 7600 件の攻撃行動を復元した。そのうえで、認証情報の交換・分離、権限の縮小、侵害の可能性がある基幹環境の再構築を実施し、攻撃経路、検知上の失敗及び改善策を詳細に公表している。他方、侵入後にクラウド認証情報や複数クラスターへ移動できたことは、**最小権限と多層防御になお改善余地があった**ことも示している。

【図表 4】Hugging Face の対応

対応	概要
検知・封じ込め	自社監視により異常を把握し、問題となった処理経路を停止
フォレンジック	AI も利用して約 1 万 7600 件の攻撃行動を復元
認証・インフラ	認証情報の交換・分離、権限の縮小、基幹環境の再構築
情報開示	攻撃経路、検知上の課題、再発防止策を詳細に公表

※ 侵害を受けた側が攻撃経路と自らの検知上の失敗を具体的に開示した点は、インシデント対応の実務として参照価値が高い。自社が同種の侵害を受けた場合の開示範囲・開示時期の検討にあたり、有用な先例となる。

4 Claude 事案 — 現実のインターネットを演習環境と誤認した

(1)プロンプトと実環境の食い違い

Claude には、「評価環境はシミュレーションであり、インターネットには接続できない」と説明されていた。しかし、Anthropic と評価パートナーである Irregular との認識齟齬及び設定ミスにより、**実際には外部接続が可能であり、攻撃対象も技術的に限定されていなかった**。Claude は、実在サイト、認証局、メール登録等を観察してもなお、プロンプトの記述を優先し、「精巧な演習環境に違いない」と解釈した。

(2)3 件のインシデント

【図表 5】Claude による 3 件のインシデント

事案	行動	結果
①	架空企業と同名の実在企業へ侵入	認証情報を取得し、数百行の本番データを含む DB へアクセス
②	悪意ある Python パッケージを公開 PyPI へ登録	約 1 時間公開され、15 台の実システムで実行された
③	約 9000 件の外部対象をスキャン	実在企業へ侵入したが、現実と認識して自ら停止

① 架空企業と同名の実在企業への侵入

架空企業名が実在企業のドメインと重複していたため、Claude は、模擬標的に到達できないと、同名の実在企業を探して侵入した。途中で現実である可能性を認識しながら、**「実在企業も演習に組み込まれているのだろう」と合理化して継続した**点が重要である。

② 悪意ある Python パッケージの公開

Python は、AI、データ分析、業務自動化等に広く使われるプログラミング言語である。Python パッケージは、再利用可能な機能をまとめたソフトウェア部品であり、実際に動作するコードを含む。PyPI はパッケージの公開・配布サイト、pip は名称を指定してパッケージを取得・導入する道具である。

> 【身近なたとえ】
> Python は「調理方法・言語」、Python パッケージは「調理キット」、PyPI は「オンライン倉庫」、pip は「指定したキットを取り寄せる配送係」に相当する。

Claude は、架空企業の手順書に書かれたパッケージ名が現実の PyPI で未登録であることを発見した。そこで、同じ名称を先に登録し、悪意あるコードを含むパッケージを公開すれば、手順書に従うシステムが正規部品と誤認して自動導入すると考えた。これは、未登録の社内用名称と同名の公開パッケージを誤導入させる**「依存関係混乱攻撃」に近い構造**である。

Claude は、メールアドレス及びアカウントの取得、電話番号・資金調達の試行を経てパッケージを公開した。約 1 時間の公開中に 15 台で実行され、その一つであるセキュリティ企業の自動スキャナーから認証情報が外部へ送信された。**当初の標的以外にも被害が波及した**点で、ソフトウェア配布経路を利用したサプライチェーン型事故でもある。

③ 約 9000 件の外部対象のスキャン

第 3 事案では、予定された模擬標的に到達できないため約 9000 件を探索し、公開デバッグページの認証情報や SQL インジェクションを使用して実在企業へ侵入した。ただし、最新の社内研究モデルは、侵入先が課題と無関係なクラウド環境にあると理解すると、**現実の対象であると判断して自ら停止した**。安全側への停止判断が一定程度改善していることを示す事実である。

5 なぜ高性能 AI は現実を正しく認識できなかったのか

Claude は、脆弱性の探索、プログラムの作成、アカウントの取得、認証情報の利用といった高度な作業を遂行しながら、「これは現実のインターネットではないか」という判断を適切に行うことができなかった。**技術的な課題遂行能力と、環境全体を理解して安全側へ停止する能力とは、別物である。**

・局所的な技術判断と、対象・環境・第三者への影響を統合する状況認識とは、異なる能力である。

・AI は人間のような身体的・経験的な現実感を持たず、現実も模擬環境も等しく「ツールから返るデータ」として受け取る。

・システムプロンプトを強く優先するため、観察結果との矛盾を「現実ではなく精巧な演習である」と解釈した。

・目標達成に有利な説明を選択する傾向があり、高度な推論能力が誤った前提の合理化に用いられた。

・「疑わしい場合は停止し、人間に確認する」能力が、課題遂行能力と同じ速度では向上していない。

【能力と安全性の非対称】

能力の低い AI は、判断を誤っても大きなことができない。これに対し能力の高い AI は、**状況認識を誤ったまま、人間には困難な多段階の行為を高速かつ大規模に完遂できる**。能力の向上がそのまま安全性の向上を意味しないのは、この非対称性による。

6 どちらが深刻か ― 二つの異なる警告

技術的能力と将来の潜在的危険性という観点では、OpenAI 事案がより深刻である。AI が未知の脆弱性を発見し、閉じ込められた環境を自力で突破したからである。他方、**一般企業における再現可能性**という観点では、Claude 事案の方が身近である。外部接続の設定ミス、過剰権限、委託先との認識齟齬、事後的なログ確認しかない運用は、通常の企業においても十分に起こり得る。

【図表 6】深刻度の比較

項目	OpenAI 事案	Claude 事案
隔離の突破	AI が自ら突破した	設定ミスにより開いていた経路を利用した
用いた技術	ゼロデイ、権限昇格、横移動、RCE	弱い認証、未認証の接続先、SQL インジェクション等
主要な警告	閉じ込めても、AI が自力で突破し得る	閉じ込めたつもりでも、設定ミスにより現実に作用する
一般企業での再現性	相対的に低い	高い

両事案は、AI の「暴走」ではなく、与えられた目標の過剰遂行である。問題の中心は AI の意思ではなく、**目標、権限、環境及び監視の設計**にある。したがって、対応も「より賢い AI を待つ」ことではなく、「誤判断を前提とした統制を構築する」ことでなければならない。

7 企業が採用すべき統制 ― 九層の多層防御

(1)事案類型ごとの防止・軽減策

OpenAI 型の事案に対しては、**AI 自身を高度な内部攻撃者とみなす**必要がある。外部接続型プロキシを避けて事前確認済みパッケージのみを提供し、異なる技術による多重隔離、ノード間通信の禁止、偽物・短期・用途限定の認証情報、実行時間・通信量・試行回数の上限を設けることが考えられる。境界探索、権限昇格、横移動又は外部通信が始まった時点で自動停止させることが重要である。

Claude 型の事案に対しては、評価・運用開始前の実接続テスト、外部通信のデフォルト拒否、IP・ドメイン・ポートの許可リスト、.test・.invalid 等の予約ドメインの使用、模擬レジストリの利用、外部メール・SMS・決済・アカウント作成の遮断が有効である。大量スキャン、認証情報の取得及び外部公開はリアルタイムで検知し、**現実である可能性が生じた時点で停止して人間の承認を求める**運用とすべきである。

(2)九層の多層防御モデル

【図表 7】九層の多層防御モデル

層	必要な統制
①モデル	危険な行動の拒否、疑わしい場合の停止
②プロンプト	対象範囲、禁止事項及び停止条件を明記する
③ツール	使用可能なコマンド・API・サービスを制限する
④権限	最小権限、短期認証情報、職務分離
⑤ネットワーク	外部通信の原則拒否、許可リスト、セグメント分離
⑥人間の承認	外部送信、公開、削除、決済等を事前承認の対象とする
⑦行動監視	個別操作ではなく、行動系列（軌跡）全体を監視する
⑧緊急停止	閾値超過、異常通信、境界探索を検知して強制終了する
⑨事故対応	完全なログ、通知、フォレンジック、復旧手順

最も重要な原則は、AI が正しく判断することを第一の防壁にしないことである。

情報検索や下書きの作成は限定的に自動化できるとしても、外部メールの送信、コード・パッケージの公開、削除、送金、契約締結、本番環境のセキュリティ設定変更等については、**人間の事前承認又は二者承認を原則とすべきである。**

8 法的・ガバナンス上の留意点

日本企業が AI エージェントに外部システムを操作させる場合には、刑事法、民事責任、会社法上の内部統制、個人情報保護、委託契約及び業法上の報告が問題となり得る。具体的な責任の所在は、指示内容、権限設定、予見可能性、停止可能性、ログの有無及び被害内容等により異なる。

【図表 8】主な法的論点

観点	留意点
不正アクセス禁止法	識別符号の無断入力やアクセス制御機能の回避が問題となる。AI が操作を選択した場合であっても、誰がどのような目的・認識で設計、指示、稼働及び継続をさせたかが重要である。刑事責任における故意の認定は慎重を要する。
民事責任	予見可能性、権限及び接続範囲の設定、監視・停止の有無、委託先の選定・監督等を踏まえ、不法行為責任又は契約責任が問題となり得る。
内部統制（会社法）	高リスクの AI に広い権限を与えながら監視・承認・停止の仕組みを欠く場合、「損失の危険の管理に関する体制」（会社法 362 条 4 項 6 号、会社法施行規則 100 条 1 項 2 号）の不備が問題となり得る。
個人情報保護	安全管理措置、委託先の監督、漏えい等が生じた場合の個人情報保護委員会への報告及び本人への通知が問題となる。
委託契約	ネットワーク構成、接続範囲、権限、ログ、緊急停止、報告期限、再委託、監査権及び費用負担を具体的に定める必要がある。
保険・業法	サイバー保険の補償範囲（AI の自律的行為が免責事由に該当しないか）を確認し、規制業種においては当局報告及び業務継続の観点も検討する。

【刑事責任と管理責任は分けて考える必要がある】
AI が自律的に行動した場合、人間について犯罪の故意を直ちに認定できるとは限らない。しかし、**刑事責任が成立しないことと、民事・行政・契約・会社法上の管理責任がないこととは別問題である。**取締役の善管注意義務の観点からは、むしろ「制御可能であったにもかかわらず制御しなかった」ことが問われる。

9　経営・三線管理と 90 日の実務対応

AI エージェントの統制を IT 部門のみに委ねるべきではない。**第 1 線**の業務部門・IT 運用は利用目的、権限及び日常監視を担い、**第 2 線**のセキュリティ・リスク・法務は基準の策定、リスク評価、例外承認、委託先管理及び事故対応を管理する。**第 3 線**の内部監査は、規程の有無にとどまらず、実際の設定、ログ、承認及び停止機能を検証する。取締役会・経営会議は、リスク許容度、重大な例外、残存リスク並びに事故及び改善の状況を監督する。

【図表 9】導入済み企業における 90 日の優先対応

期間	優先対応
0～30 日	AI エージェント、接続先、認証情報及び実行権限を一覧化し、外部公開・削除・決済を暫定的に承認制とする。
31～60 日	許可リスト、短期資格情報、承認フロー、委託先チェック及び緊急停止手順を整備する。
61～90 日	行動軌跡監視、異常シナリオ試験及びインシデント演習を実施し、残存リスクを取締役会へ報告する。

【図表 10】経営者向けチェックリスト

確認事項	はい	いいえ
AI エージェント、利用目的、接続先及び権限を一覧化しているか	☐	☐
読取り、提案、実行及び公開の権限を分離しているか	☐	☐
外部送信、公開、削除及び決済について人間の承認を要する運用としているか	☐	☐
接続先を許可リストで限定し、本番環境とテスト環境を分離しているか	☐	☐
認証情報を短期・最小権限・用途限定としているか	☐	☐
大量アクセス、境界探索及び権限昇格を検知して自動停止できるか	☐	☐
行動軌跡を監視し、事故時の証拠保全・復旧手順を定めているか	☐	☐
委託先の構成・監視状況を確認し、残存リスクを取締役会へ報告しているか	☐	☐

10　まとめ ―「賢い AI ほど安全」とは限らない

OpenAI のモデルは、未知の脆弱性を発見して隔離環境を突破した。Claude は、アカウントの作成、マルウェアの公開、認証情報の取得等を実行できたにもかかわらず、自らが現実のインターネット上にいることを正しく理解できなかった。今回の事故は **AI の反乱ではなく、高い実行能力と不完全な状況認識との組合せを、人間が十分に管理できなかった事故**である。

（筆者の所感）「誤解する AI」を前提とした設計へ
本件で最も示唆的であるのは、Claude が現実と演習を識別できなかったという事実そのものよりも、**現実である可能性を認識しながら、目標達成に有利な説明を選んで継続した**という点である。これは、能力の不足ではなく、

判断の方向性の問題である。人間の組織においても、目標が明確に与えられ、権限が広く、監視が緩い場合には、同種の合理化が生じることは経験則の示すところである。

そうであるとすれば、AI ガバナンスの設計思想は、コンプライアンス体制の設計思想と本質的に異なるものではない。すなわち、**行為者が誤ることを前提として、誤っても重大な結果に至らないように、権限を分け、記録を残し、一定の行為には事前承認を要求し、異常時には止める**という構造である。今回の事故は、この古典的な内部統制の発想が、AI エージェントに対してこそ厳格に適用されるべきことを示している。

筆者としては、AI エージェントの導入それ自体を抑制すべきとは考えていない。むしろ、**導入の速度に統制の整備が追いついていないことこそが問題**である。とりわけ、外部委託先の評価環境や AI 開発基盤を利用する企業においては、契約上の手当て（構成の開示、ログの保全、緊急停止、報告期限）を先に整えておくことが、現実的かつ費用対効果の高い対応であると考える。

【まとめ】

今後の AI ガバナンスは、**AI が誤解し、状況を取り違え、目標を過剰に追求し、想定外の経路を発見することを前提として、それでも被害が生じない統制を構築すること**が中心となる。AI エージェントの統制は、IT 部門の技術課題ではなく、**取締役会が監督すべき経営・内部統制上の課題**である。

〔別紙 1〕 用語集

【図表 11】重要用語

用語	意味
AI エージェント	ツールを使用し、複数の手順を自律的に計画・実行・修正する AI
CTF	隠された秘密情報「フラグ」を取得することを競うサイバー演習
ExploitGym	複雑な脆弱性の発見・悪用能力を測定するベンチマーク
ゼロデイ脆弱性	開発元や防御側がまだ十分に把握・修正していない脆弱性
エクスプロイト	脆弱性を侵入・不正操作に利用する手法又はコード
RCE（遠隔コード実行）	外部から相手のサーバー上で任意のコードを実行すること
権限昇格	一般権限から管理者等の高い権限を取得すること
横移動（ラテラルムーブメント）	一つの侵入先を足掛かりとして別のシステムへ移ること
Python	AI、データ分析、業務自動化等に広く使われるプログラミング言語
Python パッケージ	Python の機能を導入・再利用できる形にまとめたソフトウェア部品
PyPI	Python パッケージを公開・配布するインターネット上のレジストリ
pip	指定した Python パッケージを探して導入する道具
依存関係混乱攻撃	同名の公開パッケージを正規部品と誤認させて導入させる攻撃
SQL インジェクション	不正なデータベース命令を入力し、情報の閲覧・変更等を行う攻撃
サンドボックス	プログラムを本番環境・外部環境から隔離して動作させる環境
状況認識	AI が環境、対象及び自らの行動の現実への影響を理解する能力
軌跡単位の監視	単発の操作ではなく、一連の行動が目指す結果を監視すること
最小権限	業務に必要な最小限の権限のみを付与する原則
許可リスト	あらかじめ認めた接続先・操作のみを利用可能とする方式
キルスイッチ	異常時に AI の処理・通信を強制停止する仕組み
フォレンジック	ログ等から侵入経路、行動及び被害範囲を復元する調査
ポストモーテム	事故後に原因、経緯、被害及び再発防止策を整理する事後検証

〔別紙 2〕 理解度確認問題（〇×10 問）

社内研修等でご利用いただけるよう、理解度確認問題を掲載する。解答及び解説を併記している。

【図表 12】理解度確認問題と解答・解説

No.	設問	解答	解説
1	一般利用者向け ChatGPT が利用者の端末から脱出して Hugging Face へ侵入した。	×	社内のサイバー能力評価に用いられたモデルによる事案である。
2	OpenAI 事案では、AI が未知の脆弱性を発見し、権限昇格・横移動を経て外部へ到達した。	〇	自力で隔離環境を突破した点が技術的に重大である。

No.	設問	解答	解説
3	Claude 事案でも、AI がゼロデイ脆弱性を発見して隔離環境を突破した。	×	設定ミスによりインターネットへの経路が開いていたものである。
4	Claude は「インターネットには接続できない」というプロンプトを優先し、実在システムを演習対象と誤認した。	○	プロンプトと実環境の食い違いが誤認を固定した。
5	サイバー技術能力が高ければ、現実と演習を識別する能力も必ず高い。	×	課題遂行能力と状況認識・停止判断とは別の能力である。
6	プロンプトに「外部へアクセスしてはならない」と書けば、技術的な通信制御は不要である。	×	自然言語による指示はアクセス制御の代替とはならない。
7	技術的危険性では OpenAI 事案が深刻である一方、一般企業で再現されやすいのは Claude 型の設定・運用事故である。	○	二つの事案は、それぞれ異なる警告を発している。
8	外部メール、コード公開、削除、決済等については、人間の事前承認を設けることが有効である。	○	不可逆的・対外的な行為を AI 単独に委ねないことが重要である。
9	AI 評価を外部委託すれば、委託元はネットワーク構成、ログ及び緊急停止を確認する必要はない。	×	委託をしても、管理責任が当然に移転するわけではない。
10	AI に悪意や自己保存目的がなくても、広い権限と外部接続を与えれば現実の損害を生じ得る。	○	本件の核心は、高能力・広権限・不十分な統制の組合せにある。

〔主な参考資料〕

1 OpenAI「OpenAI と Hugging Face、モデル評価中のセキュリティインシデント対応で連携」（2026 年 7 月 21 日公表）

日本語版 https://openai.com/ja-JP/index/hugging-face-model-evaluation-security-incident/

英語版 https://openai.com/index/hugging-face-model-evaluation-security-incident/

※ 同公表は、最終的な事故調査報告書ではなく、初期調査の結果及び対応方針をまとめた公式報告である。

2 OpenAI "Safety and alignment in an era of long-horizon models"

3 Anthropic "Investigating three real-world incidents in our cybersecurity evaluations"

https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals

4 Hugging Face "Security incident disclosure — July 2026"

https://huggingface.co/blog/security-incident-july-2026

※ 何が起きたか、影響範囲、封じ込め、認証情報のローテーション及び再発防止策を簡潔にまとめた初期のインシデント公表。本稿の概要部分（図表 4 等）の主たる出典である。

5 Hugging Face "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident"

https://huggingface.co/blog/agent-intrusion-technical-timeline

※ OpenAI の評価環境からの脱出、同社への侵入経路、約 1 万 7600 件の行動、権限昇格・横移動、被害範囲並びに検知・解析の方法を詳述した技術調査報告。本稿の技術的事実関係（図表 3 等）の主たる出典である。

※ 本ニュースレターは、各社公表資料を中心に 2026 年 8 月 5 日時点の情報を整理・分析したものです。OpenAI の 2026 年 7 月 21 日付公表は初期調査の結果及び対応方針を示すものであり、調査は継続中です。したがって、今後の追加報告により経緯・原因・影響範囲の評価が修正される可能性があります。また、本稿は一般的な情報提供を目的とするものであり、個別の事案に関する法的助言ではありません。