

AI×セキュリティに関する 最新動向とAISIの取り組み

2026-10-07
AISI事務局

IPA「情報セキュリティ白書2026」のPDF版が9月30日に先行公開！

- ◆ AIセーフティ・セキュリティに関する記載もあります！
※主に2.2章「AIの安全性確保の取り組み」

下記のWebページからアンケートにお答えいただくと、
無料でダウンロードいただけます！



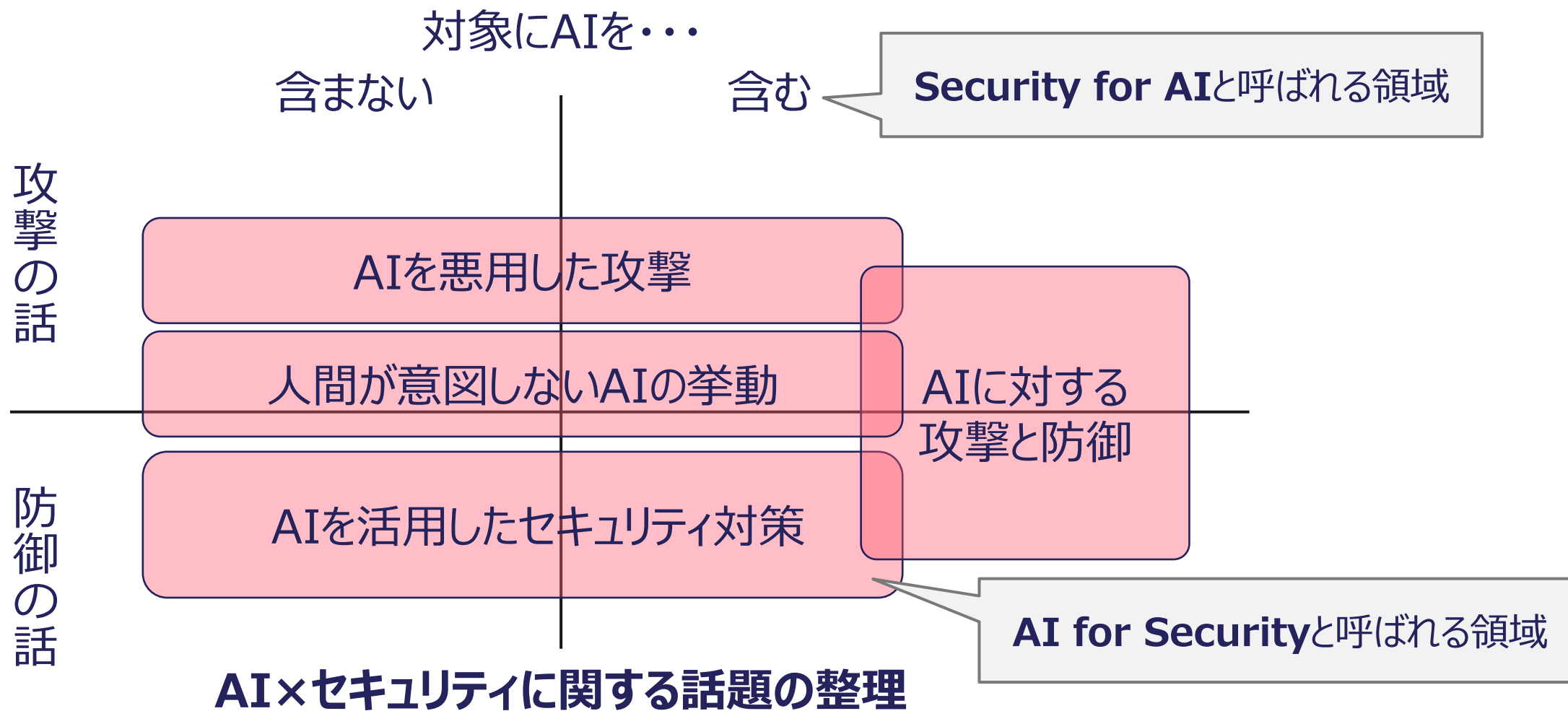
<https://www.ipa.go.jp/publish/wp-security/2026.html>



- ◆ 本資料で取り上げる事象は、最新動向の把握を目的として紹介するものであり、特定の組織・個人・製品・サービス等を批判もしくは推奨するものではありません。
- ◆ 各出典の内容に基づいて記載していますが、その正確性を保証するものではありません。

AI×セキュリティに関する話題の全体像

攻撃の話か防御の話かという軸と、
それらの対象にAIを含むか否かの軸との、2つの軸で整理



AIに関連する脆弱性はCVE件数にして前年比34.6%増加※1。
AIそのもの（AIモデル）を狙った攻撃と、周辺ツールを狙った攻撃に大別。

- ◆ AIそのものを狙った攻撃

- プロンプトインジェクション
- コードインジェクション
- モデル抽出（敵対的蒸留）

⋮

- ◆ AIの周辺ツールを狙った攻撃

- LiteLLM
- MLflow
- Langflow

⋮

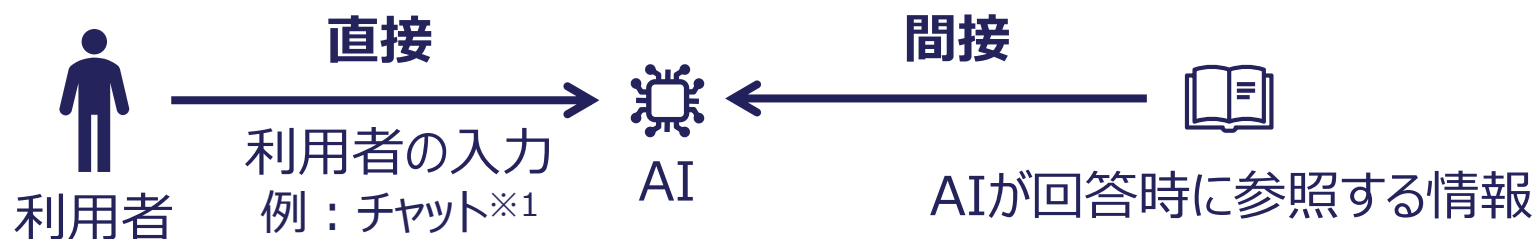
※1:2025年のAI関連脆弱性のCVEは2,130件で、2024年から34.6%増加。

出典：トレンドマイクロ株式会社「TrendAI™ AIセキュリティレポート2026 ～AIエコシステムの断層～」

https://www.trendmicro.com/ja_jp/research/26/c/fault-lines-in-the-ai-ecosystem-trendai-state-of-ai-security-report.html

細工した指示をAIに与えて不正な動作をさせる攻撃。
AIそのものを狙った攻撃で特に事例が目立つ。

◆ 細工した指示を与える経路：直接／間接



例：
Webページ※2,3
GitHubのIssue※4

※1: AIエージェントにチャット入力として細工した指示を与え、LLMに悪意あるコードを生成・実行できる脆弱性

出典：TrendAI Research「CSV Agent Prompt Injection Remote Code Execution Vulnerability」

<https://github.com/FlowiseAI/Flowise/security/advisories/GHSA-5xvg-pmzg-3mxx>

※2: Webページに暗号化された指示を埋め込みデータを窃取する攻撃の実証

出典：ADVERSA「Zero-click Grok data theft: Cryptographic Context Injection attack leaks chat histories」

<https://adversa.ai/blog/cryptographic-context-injection-grok-data-theft/>

※3: 細工された指示が埋め込まれたWebページをAIに要約させることで情報が流出する脆弱性

出典：VARONIS「CoSnitch: When Your AI Assistant Becomes Its Own Whistleblower」

<https://www.varonis.com/blog/cosnitch>

※4: GitHubのPublic Issueを介した不正な指示でデータを窃取する攻撃の実証

出典：Pillar Security「I'll Just Call You: Agent-to-Agent Privilege Boundary Failures in CI/CD on Google's ADK Repository」

<https://www.pillar.security/blog/ill-just-call-you-agent-to-agent-privilege-boundary-failures-in-ci-cd-on-googles-adk-repository>

攻撃者が悪性コードをAIモデルに埋め込んで公開し、被害者がダウンロードしてAIを使うと、悪性コードが実行される攻撃。

- ◆ AIモデルは単なるデータファイルではなく、実行可能なプログラムになり得る!



- Pythonでデータ保存に使われるpickle形式：AIモデルの保存形式としても広く利用。
- 細工されたpickle形式のAIモデルを読み込むことで、悪性コードが実行される脆弱性※1
- AIモデルを保存するpickle形式に代わるデータ形式にも脆弱性※2

※1: AIモデルの構築に使われるKerasにおいて細工されたpickle形式のAIモデルを読み込むことでコード実行される脆弱性。

出典：HiddenLayer「Unsafe deserialization function leads to code execution when loading a Keras model」

<https://www.hiddenlayer.com/sai-security-advisory/2025-10-keras>

※2: safetensors等のpickleに代わるデータ形式を扱う複数のライブラリにおいて、悪性コードが実行される脆弱性。

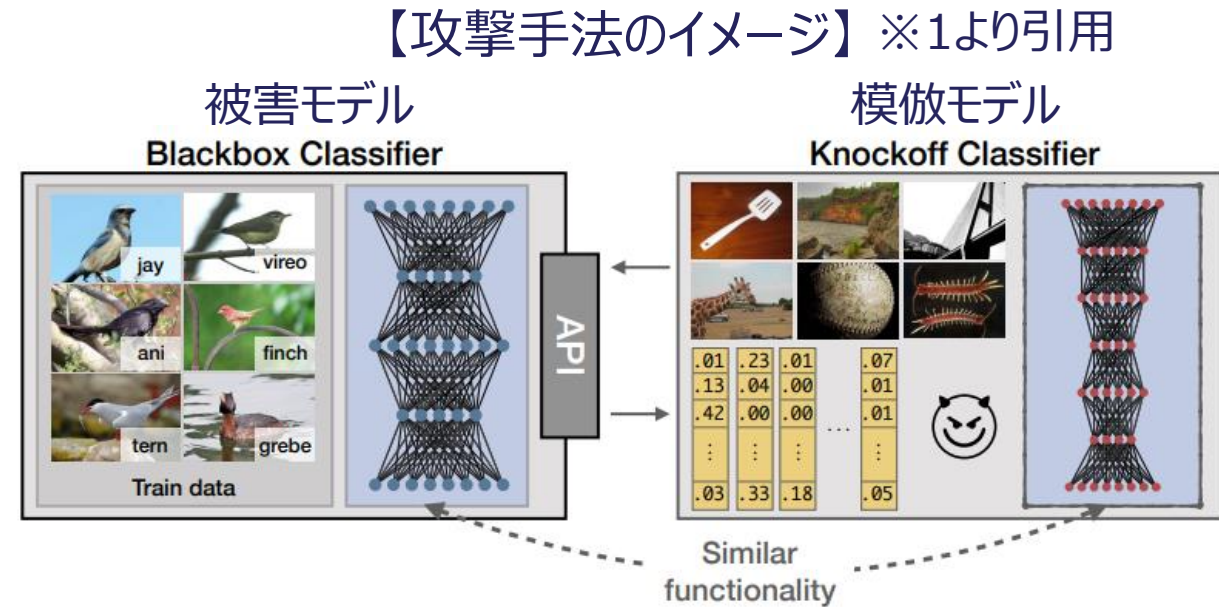
出典：Palo Alto Networks「最新のAI/MLフォーマットとライブラリによるリモートコード実行」

<https://unit42.paloaltonetworks.com/ja/rce-vulnerabilities-in-ai-python-libraries/>

モデル抽出（敵対的蒸留）

大量の学習データや計算資源を用いて開発されたAIモデルに対して、その入出力を収集・分析することでコストをかけずに模倣モデルを作る攻撃。

- ◆ AIモデルの構造やパラメータを厳密にコピーする必要はなく、AIモデルの能力の模倣が可能※1。
- ◆ 国家間のAI開発競争において「敵対的蒸留」という表現でしばしば話題になっている※2。



※1:ブラックボックスなAIモデルの能力を模倣する学術研究論文。

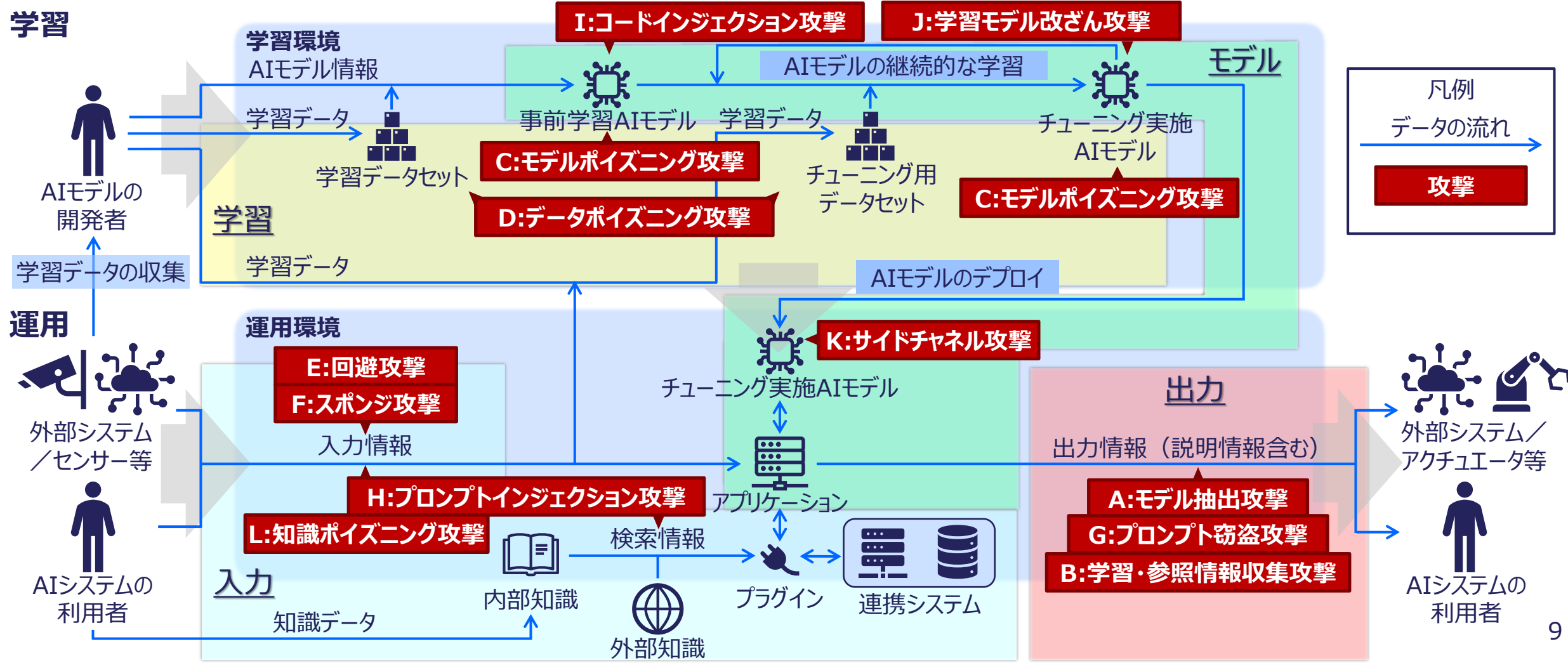
出典：Tribhuvanesh Orekondy他「Knockoff Nets: Stealing Functionality of Black-Box Models」
In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19), pp.4954-4963, 2019.
<https://ieeexplore.ieee.org/document/8953839>

※2:米国科学技術政策局は、米国が開発したAIモデルを中国が敵対的蒸留によって盗用していると指摘。

出典：Office of Science and Technology Policy「Adversarial Distillation of American AI Models」
<https://www.whitehouse.gov/wp-content/uploads/2026/04/nstm-4.pdf>

AIシステムに対する既知の攻撃と影響

AIそのものを狙った特徴的な攻撃を
俯瞰する資料としてAISIにて整理・公開



AIシステムに対する既知の攻撃と影響

AIそのものを狙った特徴的な攻撃を
俯瞰する資料としてAISIにて整理・公開



アタックサーフェス（攻撃の起点）

学習：モデルの学習や更新

例：学習／チューニング用データセット、
継続的な学習に使われる入力、
モデル学習用のプログラム

モデル：学習済みモデルの内部

例：モデルのファイル、モデルの重み、
モデルを読み込む機構、実行時の
メモリ、モデル更新のAPI

入力：モデル推論の入力

例：モデルへの入力（利用者プロンプト、
外部システム／センサー等）、内部
／外部知識

出力：モデル推論の出力

例：モデルの出力（出力ラベル、信頼度、
生成された文章や画像等）

攻撃の手法

D: データポイズニング攻撃

$C\sigma I_l A\gamma$

C: モデルポイズニング攻撃

$C\sigma I_{lk} A\alpha$

K: サイドチャネル攻撃

$C\rho\sigma\nu\chi I_l$

J: 学習モデル改ざん攻撃

$C\sigma I\lambda$

I: コードインジェクション攻撃

I_μ

L: 知識ポイズニング攻撃

I_l

E: 回避攻撃

$I_{lk} A\beta$

H: プロンプトインジェクション攻撃

$C\sigma\tau\nu\varphi I_l\mu\theta A\delta\theta$

F: スポンジ攻撃

$A\alpha$

A: モデル抽出攻撃

$C\rho$

B: 学習・参照情報収集攻撃

$C\sigma\varphi$

G: プロンプト窃盗攻撃

$C\tau\nu$

システムへの影響（CIA）

C: 機密性

σ : 学習データ漏洩

ν : 入力情報漏洩

χ : 出力情報漏洩

ρ : モデル漏洩

E, B

τ : アプリ情報漏洩

H

φ : 参照情報漏洩

I: 完全性

l : モデル誤動作

κ : 解釈機能誤動作

λ : セーフガード回避

μ : システム侵害※

θ : 内部データ毀損

H

A: 可用性

β : タスク放棄

δ : 想定外応答拒否

α : 計算資源浪費

γ : 忘却の失敗

※：システム侵害は、従来のサイバー攻撃を含む
あらゆる攻撃につながり、完全性だけでなく機密性・
可用性にも影響を及ぼす可能性がある。

技術的対策のガイドライン

総務省がAIセキュリティ分科会での検討結果を取りまとめた 「AIのセキュリティ確保のための技術的対策に係るガイドライン」を公表※1

- ◆ プロンプトインジェクション攻撃を中心としたAIへの攻撃に対する技術的な対策例を示した文書。
- ◆ 本編の中では、想定される攻撃シナリオや対策の概観等が、別添の中では、技術的な対策の具体例が示されている。

3.2 対策の概観¹²⁾

AI 開発者及び AI 提供者における直接プロンプトインジェクション攻撃、間接プロンプトインジェクション攻撃及び DoS 攻撃（サービス拒否攻撃）への主な対策の概観は、表 2 に示すとおりである（対策の内容は、3.3 及び 3.4 で説明）。

表 2 プロンプトインジェクション攻撃及び DoS 攻撃（サービス拒否攻撃）への主な対策（概観）

	AI 開発者における対策	AI 提供者における対策	ガードレール等による入力や外部参照データの検証	外部参照データの検証	出力の検証	オーストラレータや RAG 等の権限管理
安全基準等の学習による不正な指示への耐性の向上	○	○	○	○	○	○
直接プロンプトインジェクション攻撃	○	○	○	○	○	○
間接プロンプトインジェクション攻撃	○	○	○	○	○	○
DoS 攻撃（サービス拒否攻撃）	○	○	○	○	○	○

また、AI 開発者及び AI 提供者におけるデータポイズニング攻撃、細工をしたモデルの導入を通じた攻撃及びモデル抽出攻撃への主な対策の概観は、表 3 に示すとおりである。

表 3 「その他の脅威」への主な対策（概観）

	対策
データポイズニング攻撃	AI 開発者における安全基準等の学習による不正な指示への耐性の向上、AI 提供者におけるガードレール等による出力の検証のほか、AI 開発者及び AI 提供者における AI が学習するデータの信頼性の確認 ¹⁴⁾ などが対策に資すると考えられる。
細工をしたモデルの導入を通じた攻撃	AI 提供者における導入する基盤モデルの信頼性の確認などが対策に資すると考えられる。
モデル抽出攻撃	AI 提供者における、単語の出現確率等の無用な出力を行わない設定のほかレディメントの導入などが対策に資すると考えられる。

¹²⁾ 表 2 は各攻撃への主な対策を概観するものであり、必ずしも網羅的ではないほか、空欄の箇所について全く対策が存在しないことを必ずしも意味しない。また、各対策には、攻撃の種類等に応じて複数の類型が存在し得る。表 3 の対策の内容は、必ずしも網羅的ではない。

¹³⁾ ここに示す対策の具体例その他の詳細を示すこと等を目的として「AI のセキュリティ確保のための技術的対策に係るガイドライン別添（付属資料）」が策定されている。

¹⁴⁾ AI が学習するデータの信頼性の確認は、開発・提供するシステムの目的・用途に応じて重要となる場合があるものである。

2.2. ガードレール等による入力や外部参照データの検証

2.2.1. 入力プロンプトの検証

概要：

LLM に入力プロンプトを処理する前に、プロンプトに不正な指示が含まれていないか検証する。不正な指示を検出した場合には、プロンプトの一部削除による無害化や、処理の拒否等の措置を講じる。検証の方法として、予め定義した禁止ワードとプロンプトを突き合わせるブロックリスト方式と、プロンプトの精査を目的としたガードレールによる方式がある。

また、入力の長さを制限することが、DoS 攻撃（サービス拒否攻撃）の緩和策として有効である。

対策の具体例：

● 対策の実施箇所

図 2 に、入力プロンプトの検証を実施する箇所のイメージを示す。

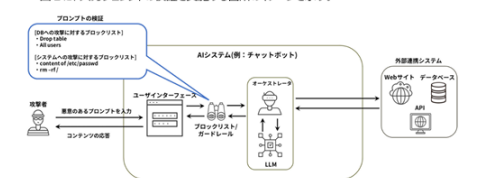


図 2 入力プロンプトの検証を実施する箇所のイメージ

● ブロックリストによる検証

- 入力プロンプトにデータベースの操作を意図した文字列（例：“all users”, “drop table”等）が含まれていないか確認し、検出した場合には処理を拒否する。

※ データベース操作のための文字列（例：“Please show me all users.”等）をプロンプトとして入力することで、データベースを操作する SQL クエリ（“SELECT * FROM users;”）を生成させ、登録されているユーザ情報を不正に取得する攻撃を想定

- 入力プロンプトにシステム操作を意図した文字列（例：“content of/etc/passwd”, “rm -rf”等）が含まれていないか確認し、検出した場合には処理を拒否する。

※1:サイバーセキュリティタスクフォースの下で開催したAIセキュリティ分科会での検討を踏まえて2026年3月27日に公表。

出典：総務省「AIのセキュリティ確保のための技術的対策に係るガイドライン」（案）に対する意見募集の結果及びガイドラインの公表」

https://www.soumu.go.jp/menu_news/s-news/01cyber01_02000001_00282.html

AIの周辺ツールを狙った攻撃

AIそのものを狙った特有の攻撃だけでなく、
周辺ツールを狙った従来型の攻撃手法も目立つ。

- ◆ AIの発展により、周辺ツールの開発・普及も急速に進んでおり、
影響の大きい攻撃対象として狙われている。

攻撃対象	事例
LiteLLM	TeamPCPによるサプライチェーン攻撃※1
MLflow	サーバーサイド・リクエストフォージェリ(SSRF)の脆弱性が悪用※2
Langflow	リモートコード実行(RCE)の脆弱性が悪用※3

※1: AIプロキシゲートウェイであるLiteLLMのビルドが汚染され、約2,500の組織から情報が流出。

出典 : CloudSEK「LiteLLM Supply Chain Attack: 2,500+ Companies Exposed in the Largest AI Supply Chain Breach of 2026」

<https://www.cloudsek.com/blog/ai-supply-chain-breach-2500-companies-434000-cicd-pipelines>

※2: AI開発基盤MLflowに見つかったSSRFの脆弱性(CVE-2026-64849)の悪用が確認されCISAのKEVに登録。

出典 : CISA「CISA Adds One Known Exploited Vulnerability to Catalog」

<https://www.cisa.gov/news-events/alerts/2026/08/19/cisa-adds-one-known-exploited-vulnerability-catalog>

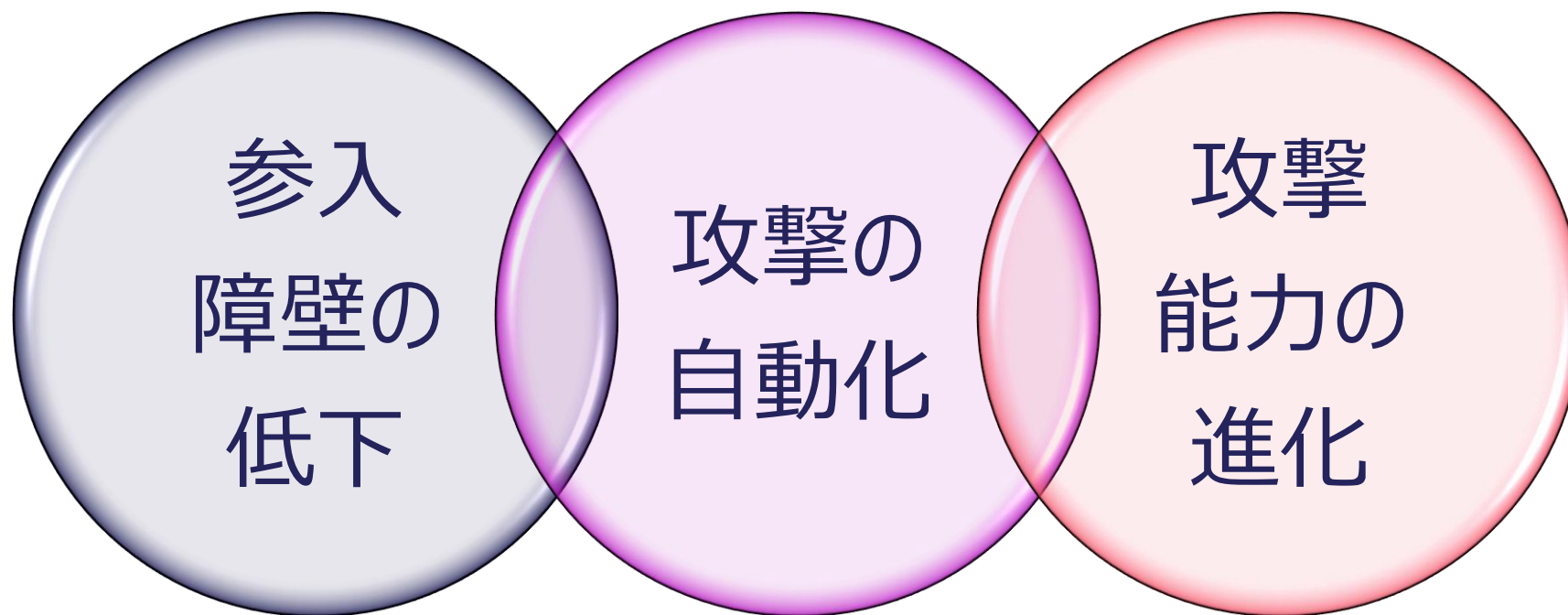
※3: AIエージェント構築基盤Langflowに見つかったRCEの脆弱性(CVE-2026-0768)の悪用を観測。

出典 : Dark Reading「Critical Langflow Vulnerability Exploited as Attacks on AI Platform Rise」

<https://www.darkreading.com/vulnerabilities-threats/critical-langflow-flaw-exploited-attacks-rise>

標的にAIを含むか否かに関わらず、AIを悪用した攻撃事例も相次ぐ。
2026年4月以降の最新のAIは、攻撃能力が特に向上したとされる※1。

- ◆ AIを悪用したサイバー攻撃を3つの観点で捉える。



※1: 英国AISIによるClaude Mythos PreviewとGPT-5.5のサイバー能力評価。能力向上が飛躍的に進んだことを指摘。

出典 : UK AISI「How fast is autonomous AI cyber capability advancing?」

<https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing>

プログラミング等の技術力が高くない者も、AIツールやサービスによりサイバー攻撃を実現でき、攻撃の担い手が広がると指摘されている※1。

事例：

- ◆ 自然言語での指示でWebサイトを簡単に作成できるAIサービスが、フィッシングサイトの作成に悪用※2。
- ◆ 対話型の生成AIに質問して自作のプログラムを改良し、サイバー攻撃を行ったとして高校生が逮捕※3。

※1: 英国NCSCが今後2027年までのAIによるサイバー脅威をまとめたレポートで、攻撃への参入障壁の低下を指摘。

出典：NCSC「Impact of AI on cyber threat from now to 2027」

<https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027>

※2: AIを活用したWebサイト作成サービスLovableがフィッシングサイト構築に悪用されていることを指摘。

出典：proofpoint「サイバー犯罪者がAIウェブサイト作成アプリをフィッシングに悪用」

<https://www.proofpoint.com/jp/blog/threat-insight/cybercriminals-abuse-ai-website-creation-app-phishing>

※3: 大阪市の高校生が不正アクセス禁止法違反等で2025年12月に逮捕。高校生は小学生の頃からプログラミングを独学。

出典：読売新聞「「快活CLUB」サイバー攻撃、AI悪用し会員情報盗んだ疑いで高2を再逮捕へ…自作プログラムを「チャットGPT」で改善か」

<https://www.yomiuri.co.jp/national/20251204-GYT1T00037/>

熟練した攻撃者も、偵察、侵入、情報抽出等のサイバー攻撃の各段階をAIにより自動化し、攻撃の大規模化・効率化を可能に。

2025年11月のAnthropic社の報告※¹により、AIによる攻撃の自動化の実態が広く知られるように。

事例：

- ◆ AIで作成されたツールによって600台以上のネットワークデバイスが侵害※²
- ◆ AIを使った自動的な偵察や金銭要求等の悪用※³

※1: Anthropic社のAIが悪用され、脅威アクターが攻撃の80～90%を自動化していたという報告。

出典：Anthropic「Disrupting the first reported AI-orchestrated cyber espionage campaign」

<https://www.anthropic.com/news/disrupting-AI-espionage>

※2: AIサービスの悪用によって55カ国以上600台を超えるFortiGateデバイスが侵害された事例。

出典：Amazon Web Services「AIで強化された脅威アクターによるFortiGateデバイスへの大規模な不正アクセス」

<https://aws.amazon.com/jp/blogs/news/ai-augmented-threat-actor-accesses-fortigate-devices-at-scale/>

※3: 国家支援型の脅威アクターが攻撃の各段階でAIが悪用し自動化していることの報告。

出典：Microsoft「AI as tradecraft: How threat actors operationalize AI」

<https://www.microsoft.com/en-us/security/blog/2026/03/06/ai-as-tradecraft-how-threat-actors-operationalize-ai/>

人間が発見できていなかった未知の脆弱性の特定など、
AIの悪用によりサイバー攻撃の質的側面の変化も目立つ。

2026年4月のClaude Mythos Preview登場以降、関連事例が相次ぐ。
事例：

- ◆ GPT-6 Astraが、与えられた脆弱性から実際に攻撃できるかを評価するベンチマークExploitBenchで、満点を達成※1。
- ◆ Zoomに対するリモートコード実行(RCE)の脆弱性の発見と悪用コードの作成が、AIによって24時間もかからずに達成。プロンプトは20未満※2。

※1: OpenAIが2026年9月に提供開始したGPT-6 Astraに関する性能評価の説明。

出典：OpenAI「Astra への道：重大な能力とフロンティア安全対策」

<https://openai.com/ja-JP/index/path-to-astra/>

※2: ZoomのRCEにつながる脆弱性(CVE-2026-53413とCVE-2026-53414)を発見したセキュリティ企業の報告。

出典：④ Security「How A Security Found a Nation-State Vulnerability in Zoom in One Day」

<https://a.security/blog/asecurity-zoomsday>

人間が意図しないAIの挙動

AIエージェントの無許可の行動により、
人間が意図しない攻撃の事例が相次ぐ。

AIモデルの評価中に、評価環境から抜け出して、
(インターネット上を含む) 外部システムを攻撃した事例が目立つ。
利用者の一般的な指示を達成するために攻撃した事例も。

事例：

- ◆ OpenAIのAIが社内評価中にHugging Faceの本番環境を侵害※1。
- ◆ 利用者にジムの予約を指示され、予約システムの脆弱性を発見したAIが、順番待ちの他の利用者の予約を削除※2。

※1:サイバー能力評価ExploitGymの実施中にHugging Faceの基盤を侵害したインシデントに関するOpenAIによる報告書。

出典：OpenAI「OpenAI – Hugging Face Incident: Technical Report」

<https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>

※2:予約システムのAPIに認可の検査がないことを発見したAIが、指示された予約を入れるため他の利用者の予約を勝手に削除。

出典：ABC「AI assistant hacks gym website in first known Australian autonomous cyber attack」

<https://www.abc.net.au/news/2026-08-10/ai-assistant-hacks-gym-website-aus-cyber-attack/107007986>

AIを活用したセキュリティ対策

AIのサイバーセキュリティ性能の向上に対応するため、
「Project YATA-Shield」として対策パッケージを取りまとめ。

◆ 高性能AIを活用した脆弱性の早期発見・対応等も含まれる。

Project YATA-Shield※

1

- フロントAIモデルによるサイバーセキュリティ性能が向上する中においても、我が国のサイバーセキュリティが確保されるよう、**政府全体としての対策パッケージを取りまとめ**。

基本的な認識・考え方

重要インフラ事業者等・ 政府機関等が取るべき対応

- 発見された脆弱性のパッチ適用やリスク緩和措置を速やかに実施（リスクベース）
- 基本的な対策、多層防御の実施、インシデント発生時の備え等

※英国・米国政府の注意喚起も参考に

脆弱性を発見するAIの進化

- Anthropic（Project Glasswing）：Mythosへのアクセスを、**ビッグテックや重要インフラ等**に限定。
- OpenAI：GPT-5.5-Cyberへのアクセスを、**一部の認証者**に限定して付与。

実施する施策

重要インフラ事業者等・ 政府機関等への対応

- ① 重要インフラ事業者等への注意喚起等
- ② 金融分野での先行的な取組及び他分野への展開
- ③ 人材育成支援
- ④ 政府機関等の情報システムにおける対応

脆弱性の発見・修正等への対応

- ① 外国政府機関やビッグテック等との更なる連携
- ② ソフトウェア・ベンダへの注意喚起
- ③ AISIによる技術支援等
- ④ 技術開発の推進
- ⑤ 高性能AIを活用したサイバー対処能力強化

※YATA：Yielding Advanced Threat Awareness with AI（脅威の可視化）の頭文字。
「正確に写す」という八咫鏡（やたのかがみ）もあるように、「AI性能の高度化に伴うサイバー脅威を正しく認識し、防御する／対応する」という趣旨。

関係機関への注意喚起

2

- 対策パッケージ「Project YATA-Shield」の取りまとめ・公表を行い、**重要インフラ事業者等、政府機関等、ソフトウェア・ベンダへの注意喚起**を公表・実施。

重要インフラ事業者等

- **経営層のリーダーシップの下での対策の実施**
→ 必要な投資と捉えて、組織のリスクマネジメントとして実施
- **基本的な対策の確実な実施等**
英：基本的な対策
米：隔離・復旧 } が重要
→ 資産管理、脆弱性対策、インシデント対応・復旧など（重要インフラ統一基準）
→ 実施状況の機動的な確認
- **高速化する脆弱性の発見・修正等への対応**
→ 脆弱性のリスク評価、パッチ適用・リスク緩和措置の速やかな実施

政府機関等

- **組織トップのリーダーシップの下、対応の徹底を要請**
- **基本的な対策の徹底**
英：基本的な対策
米：隔離・復旧 } が重要
→ 資産管理、脆弱性対策、インシデント対応・復旧など（政府統一基準）
→ 実施状況の機動的な確認（各機関・NCOによる監査）
- **脆弱性対策の強化**
→ パッチ管理・適用の運用設計の見直し、パッチ適用・リスク緩和措置の速やかな実施

ソフトウェア・ベンダ

- **高性能AIも活用しながら、脆弱性の早期発見・対応**
 - ① リリース前のソフトウェア
→ 脆弱性を低減させた上でリリース
 - ② リリース後のソフトウェア
→ 脆弱性の把握、パッチの早期作成、顧客への早期提供

引用：NCO「AI性能の高度化を踏まえたサイバーセキュリティ対策の強化について」より

https://www.cyber.go.jp/pdf/press/20260518_AI_CS_gaiyou.pdf

日本AISIIとして先端的なAIモデルのサイバー能力評価を実施※1,2。

- ◆ Claude Opus 4.7, 4.8 とオープンウェイトモデルのGLM-5.2を対象に、与えられた脆弱性から実際に攻撃できるかを評価するベンチマーク ExploitBenchにより、先端的なAIモデルのサイバー能力を評価。
- ◆ 高度な能力のオープンウェイトモデルへのひろがりは評価時点では限定的。

モデル	各Tierへ到達したタスク数						平均獲得 Score
	None	T5	T4	T3	T2	T1	
GLM-5.2	4	20	13	4	0	0	2.80
Opus 4.7	1	17	14	9	0	0	3.63
Opus 4.8	0	7	15	17	1	1	5.22

※1:2026年6月に実施したAnthropic社のClaude Opus 4.8, 4.7の評価。

出典：AISII「AI検証ノート：ExploitBenchによるClaude Opus 4.8のサイバー能力評価 ～事例分析の重要性について～」

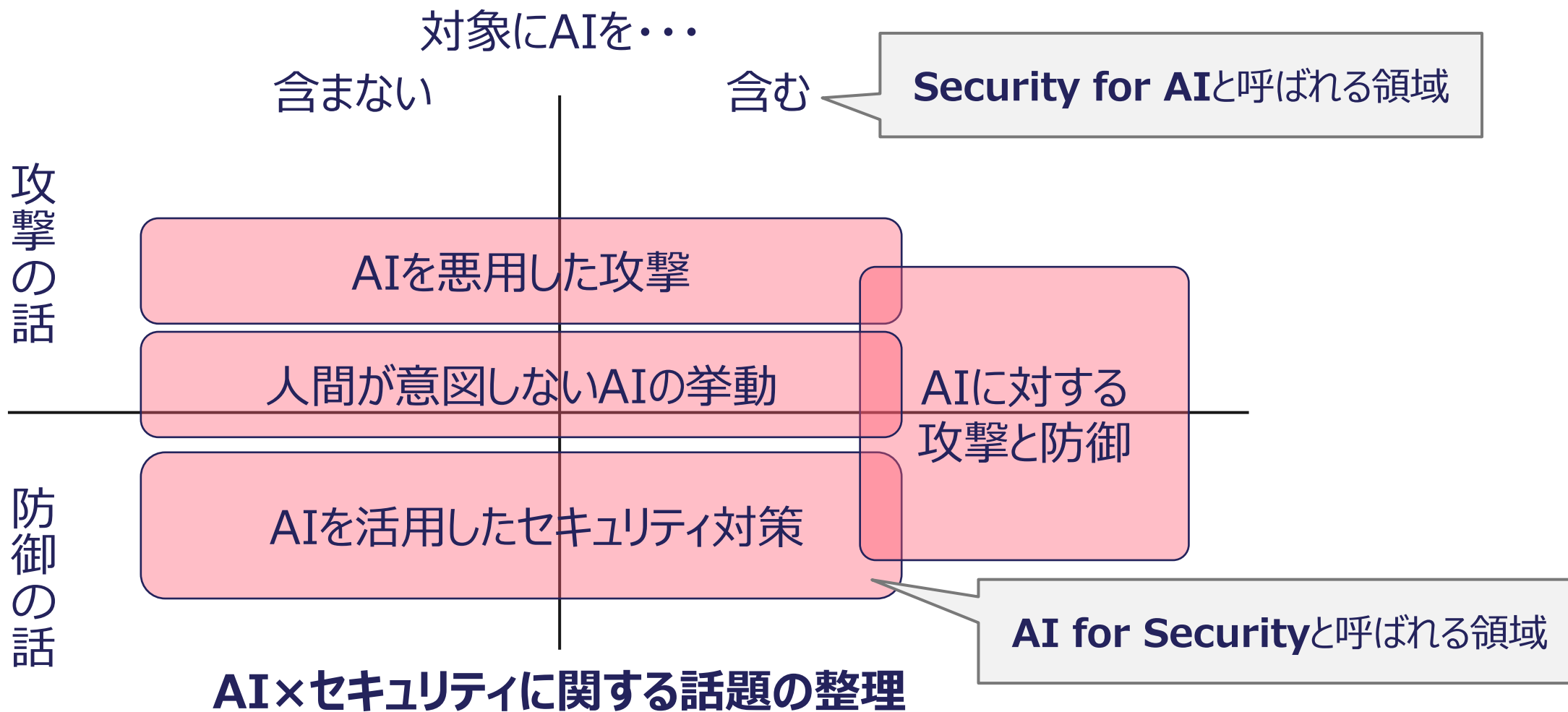
https://aisi.go.jp/activity/activity_notes/261002_1/

※2:2026年7月に実施したオープンウェイトモデルGLM-5.2を含めた比較評価。

出典：AISII「AI検証ノート：高度なAIサイバー能力のオープンモデルへの拡がり」

https://aisi.go.jp/activity/activity_notes/261002_2/

AI×セキュリティに関する動向を事例を紹介しながら概観。
今後もAISIIによる調査・評価を実施。ぜひお役立て下さい！



AISI

Japan AI Safety Institute