CISOがおさえておきたい AIセキュリティの基本

Ver.01 2026-05

JNSA CISO支援ワーキンググループ

はじめに

- AIの利用が急速に進んでいますが、AIを利用する上でのセキュリティ対策、つまり「CISOがおさえておきたいAIセキュリティ」は、あまり整理されていないように思います。
- CISO支援WG内でCISOにとってのAIセキュリティを議論をしましたが、議論の基盤となるセキュリティモデルが不確かで、有意義な議論になりませんでした。
- この資料は、「Security Days Spring 2026 Tokyo」で講演する機会を頂いた際に、WG内でのこれまでの議論を踏まえ、AI利用に係るセキュリティモデルの土台として作成したものに、「AIに係るセキュリティインシデント事例」を追加するなど、一部の内容を追記し、見直したものです。
- AIの利用環境はあまりにも急速に変化しているため、資料として公表した時点で、既に過去のものとなっている内容もあると思いますが、「CISOがおさえておきたいAIセキュリティ」を議論する土台として、変化に合わせてアップデートしていただければ幸いです。

JNSA CISO支援ワーキンググループ リーダー 高橋正和

AIのセキュリティもいろいろ

① -1 Security for AI
（AIを守る）

①-2Infrastructure &
Platform Security
（基盤セキュリティ）

② Security by AI
（AIで守る）

③ AI Safety / Responsible
AI
（安全・信頼性）

④-1 AI Misuse / Abuse
Security
（AIの悪用対策）

④-2 AIによるサイバー攻撃
（自律型サイバー攻撃など）

⑤ Secure Use of AI
（安全なAIの利用）

本講演では、主に③と⑤に
ついて取り上げます

CISOにとっての⑦ Secure Use of AI（安全なAIの利用）

CISO視点からのAI利用の懸念点

- ① データ・データ入力に係る懸念
  - 情報漏洩リスク：機密情報の意図せぬ入力・流出
  - プライバシー侵害リスク：過度なプロファイリング、同意なき学習への流用
- ② IT統制上の懸念
  - シャドーAIリスク：統制なき無許可ツールの蔓延
  - 個人情報の取り扱い：第三者提供に該当するサービスの利用
- ③ システムの応答に係る懸念　③ AI Safety / Responsible AIに相当
  - 品質リスク：ハルシネーション、誤情報に基づく経営判断
  - 法的・権利リスク：著作権、商標権の侵害
  - 倫理・コンプライアンスリスク：AIの出力における差別や偏見
- ④ AIシステムへのサイバー攻撃への懸念　① Security for AI（AIを守る）に相当
  - AI特有の攻撃リスク：プロンプトインジェクション等によるAIの悪用
  - システム基盤へのサイバー攻撃リスク：ITシステムとしての攻撃への対応

データ・データ入力に係る懸念
情報漏洩とプライバシーの懸念を整理する

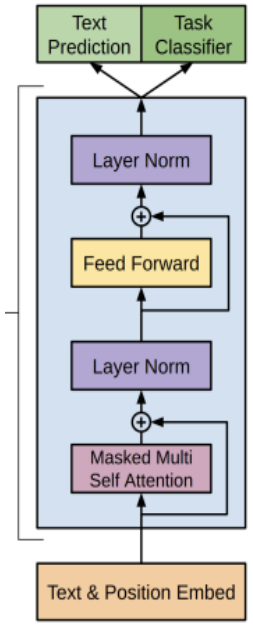
AIチャットボットの基本的な流れ

トレーニング

トレーニングデータと
推論時のデータは分離している

予測

大量の事前学習



自己教師あり
学習

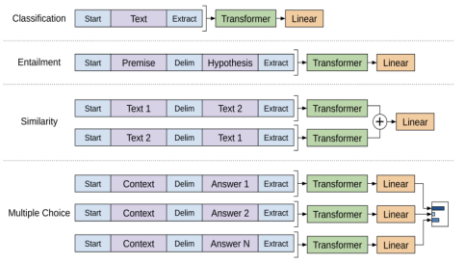
トランス
フォーマー

複数言語の学習で、各言語向にも対応できる（日本語、プログラミング言語など）

ファイン
チューニング

人間による
強化学習

RLHF: Reinforcement Learning by Human Feedback

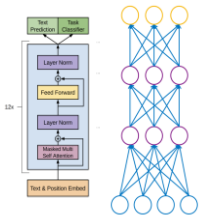


数百～数千件

学習時点までのデータ
ユーザーの入力は反映されない

チャット
ボット

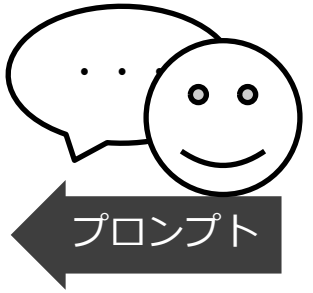
LLM



セッション
管理（Chat）

文脈内学習
In-context
Learning

プロンプトからの指示で、学習しなおすことなく、その場で様々なタスクを処理できる（Zero-shot, One-shot, Few-Shot）



自己注意機能と本文中学習

[OpenAI 2018][OpenAI 2022]
を参考に高橋が作成

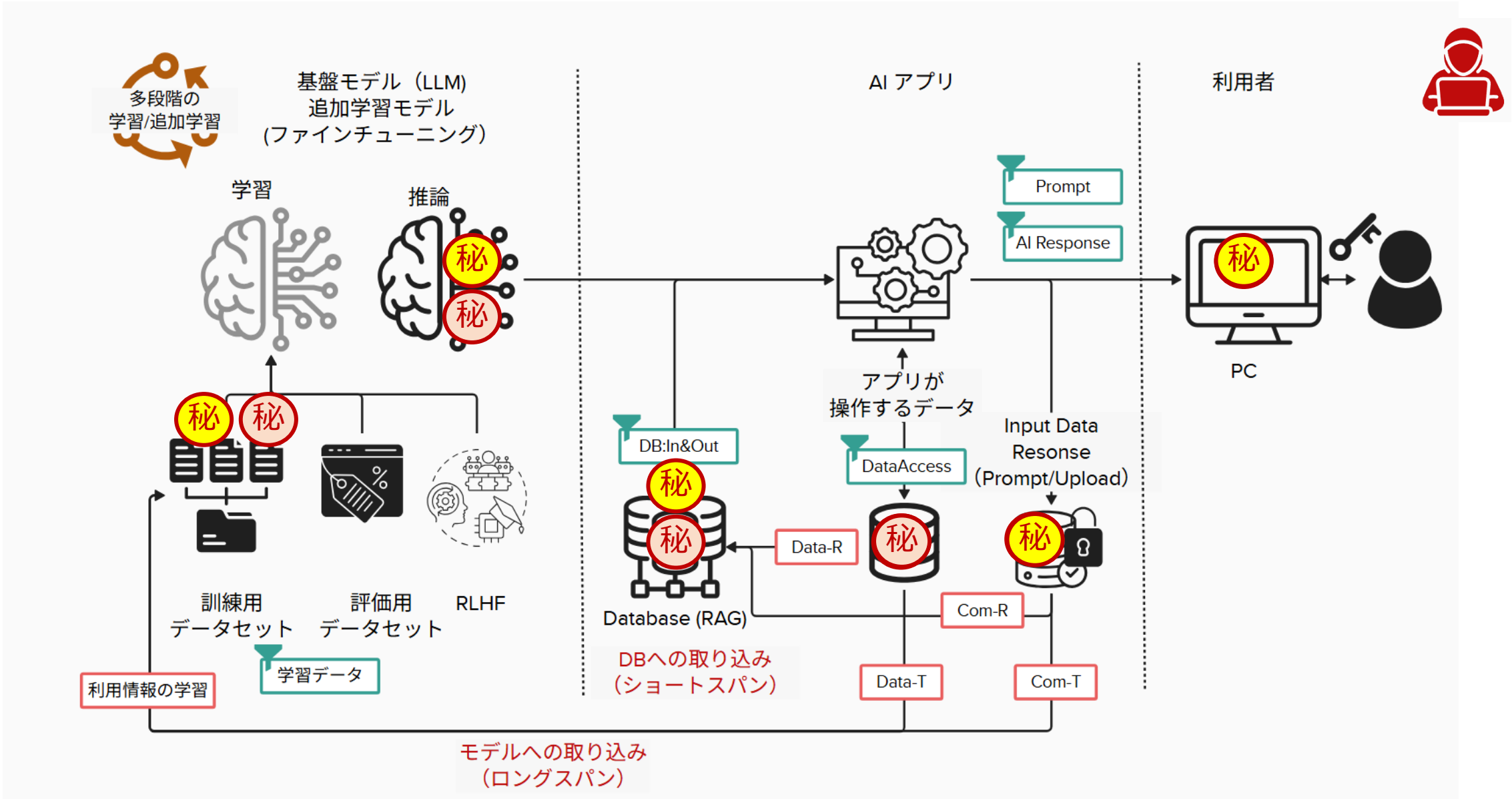
AIチャットボットが会話を記憶する仕組み

AIチャットボットと利用者のやり取りは、ユーザー側、またはチャットアプリに記録され、会話のたびにその履歴がLLMに送信されている。

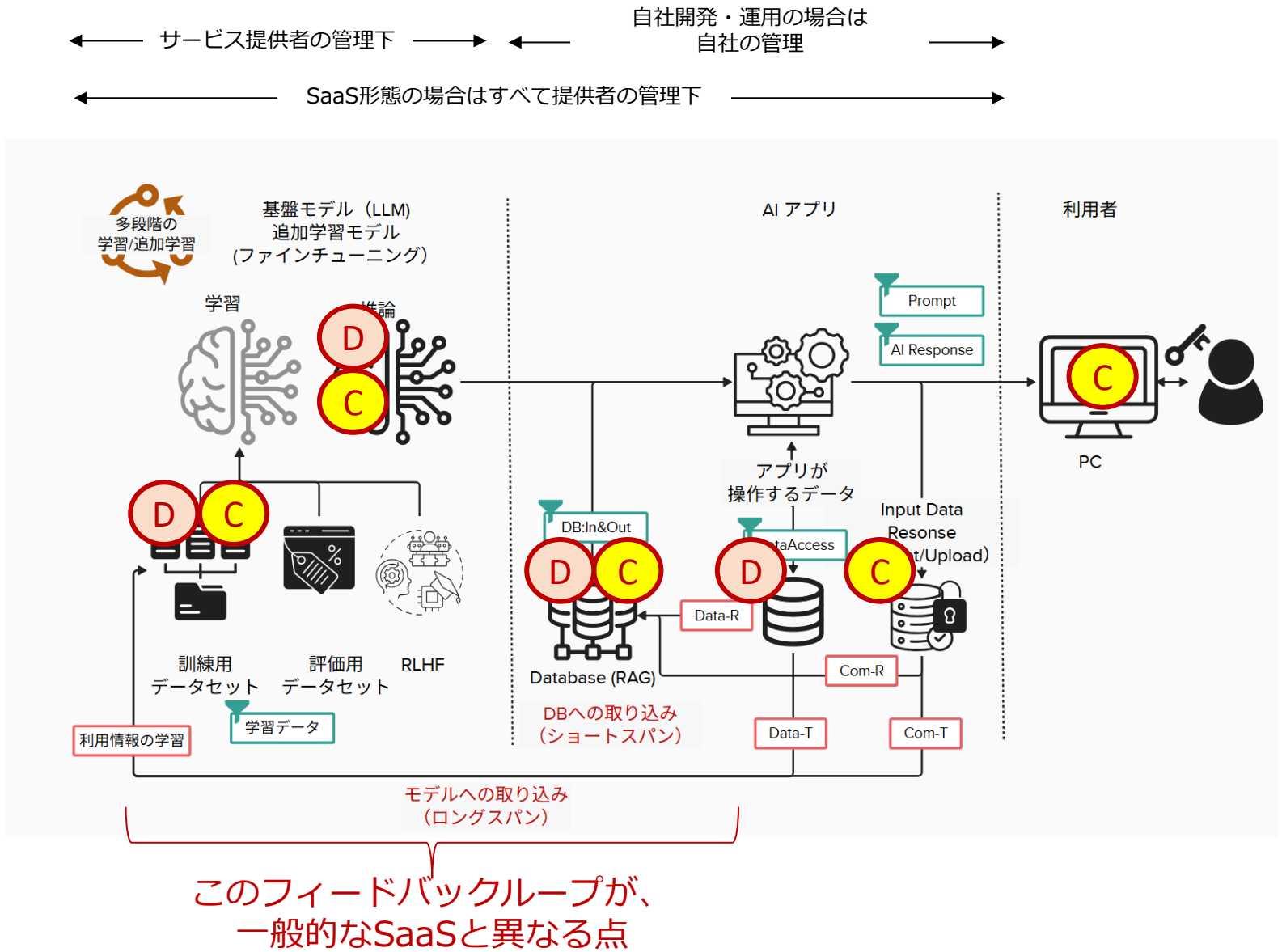

```
{
 "messages": [
  {
   "role": "user",
   "content": “example.net は、詐欺メールに使われるドメイン？"
  },
  {
   "role": "assistant",
   “content”: “結論から申し上げますと、`example.net` というドメイン自体は **exampleが所有する「本物の（正規の）ドメイン」** です。【以下省略】"
  },
  {
   "role": "user",
   "content": "このURLはどうでしょう？¥nhttps://xxxxxxz31.web.core.example.net/"
  },
{
   "role": "assistant",
   "content": "結論から申し上げますと、そのURLは極めて高い確率で「フィッシング詐欺サイト」です。絶対にアクセス（クリック）しないでください。【以下省略】"
  },
  {
   "role": "user",
   "content": "*.core. example.net/ を疑わしいドメインと考えて問題ない？"
  },
  {
   "role": "assistant",
   "content": "結論から申し上げますと、**「一律ですべて悪（アクセス禁止）とみなすのは問題がある」が、「メールのリンクとして直接送られてきた場合は、極めて疑わしい（ほぼ詐欺）と考えてよい」**というのが正しい認識になります。【以下省略】
  }
 ]
}
```


複数の端末（PC等）で、同じアカウントでチャットボットを使った場合、他の端末のやり取りが共有されるものと、共有されないものがある。

AIアプリのユーザーデータ利用モデル



AIと一般的なSaaSのデータフローの違い



このフィードバックループが、
一般的なSaaSと異なる点

ユーザー情報のソースと利用先

- Com：利用者との会話,
  Data:アプリ上のデータ,
  T: Training Data

- Com-R:　RAGに反映される会話
- Com-T:　学習データに反映される会話
- DATA-R:　RAGに反映されるアプリデータ
- DATA-T:　学習データに反映されるアプリデータ

学習データの応答境界（ Output Boundary ）
- サービス利用者全般
- 組織内（テナント）
- アカウント内に限定

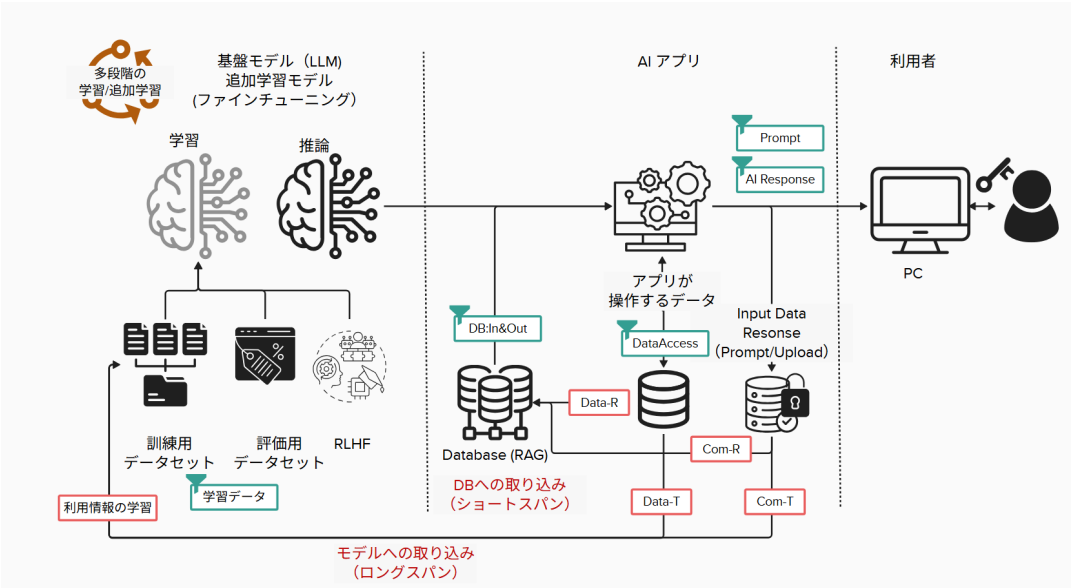
設計上の確認点（今回はスコープ外）

入力のフィルタ
- 学習データ：モデルの学習データ全般
- DB In: RAGへの入力フィルタ

出力のフィルタ
- AI Response: AIからの出力フィルタ（ガードレール）
- DB Out: RAGからの出力フィルタ

AIアプリのアクセス
- Data Access: AIアプリのアクセス権

AIアプリのユーザーデータ利用



ユーザー情報の出所と利用先
各社の公開利用規約および公式情報に基づき、情報漏洩やプライバシーの懸念を整理するために以下の表を作成した。

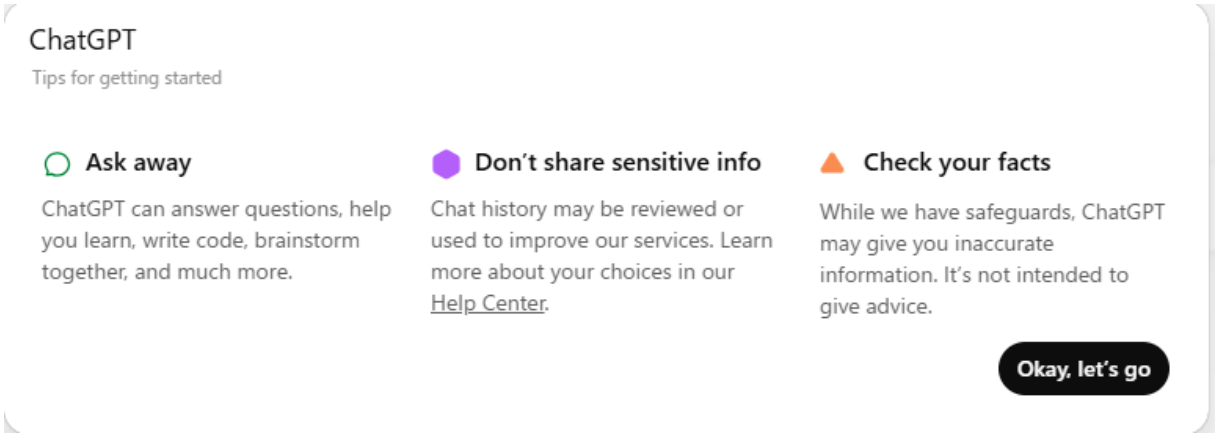
シャドーAIの利用は、入力情報が学習データに利用されることに伴って、当該AIの他の利用者に漏洩する可能性があり、広範囲な情報漏洩につながることが懸念される。

	無償版 個人用有償版Chat 全般	企業向け有償版 Chat 全般 (Gemini, ChatGPT)	Google Workspace Microsoft 365	Slack	DeepL	社内開発した AI全般（推測）
Communication	プロンプトと応答 （アップロードを含む）		直接的なAIへの指示 履歴	直接的なAIへの指示 履歴	翻訳する文と翻訳結果	顧客とのやり取り
Data	Projectsなど	Projects Gemの「知識」など	Gmail, Drive, Meet, Calendar Outlook, SharePoint, Teams	Slackの全投稿	N/A	顧客データなどの データベース
学習への利用	利用する	Enterprise/Optoutなど 利用しないサービスあり	利用しない	利用しない	無償版：利用 有償版：利用しない	学習データに 反映する可能性は低い
応答境界	全サービス利用者に 展開する可能性	テナント内に閉じる 共有も可能	Google Workspace/M365の アクセス権限	ユーザーが参加・閲覧権限 を持つチャンネルとDM	N/A	社内または顧客
RAG・DBへの取り込み	メモリ機能などによる会話の記憶		テナント専用の検索インデックスとして構築	検索インデックスの利用	N/A（ただし、無償版はモデルの学習に利用される）	会話を記録

2026年3月時点での公開情報を整理して記載。各社の最新利用規約を参照のこと

ChatGPTの例

Free版の初期画面



Enterprise版のチャット画面

③ AI Safety / Responsible AI（安全・信頼性）

AI安全性は何を「心配」しているのか？

人間中心
AIに人間が支配されない

しかし…
AIに人間が支配されないか？
人間の権利は守られるのか？

AI is too Great !
でも、もしかして
人類の危機？

主要な心配事

物理事故につながらない？ 物理的な危険（自動運転など）	安全性	基本的な権利の尊重
誰かが不利な扱いを受けない？ 民族や属性、誤ったデータ	公平性	
秘密をバラしたりしない？ メールアドレス、プロンプト	プライバシー	
ハッカーに悪用されるのでは？ ハッカーが攻撃に使うのでは？	セキュリティ	AIプロダクトへの信用・信頼の醸成
何をやっているかわからない わからないから信用できない	透明性	
言っている通りにやってる？ 本当はちゃんとやってないのでは？	アカウンタビリティ	
それって本当？嘘じゃない？ ハルシネーション、誤認識	誤った情報・動作	
危ない事に使われない？ ウィルス生成、兵器開発など	専門知識の悪用	悪用への懸念と対策
犯罪に使われない？ Deep Fake、詐欺	犯罪・犯罪行為への悪用	
私の作品パクッてない？ AIの生成物って使えるの？	権利および権利侵害	
偏見を助長してはダメでしょう？ わいせつな出力は良くないのでは？	社会倫理	

「心配」を整理・具体化する

基本的な心配事

- 人間中心
- 安全性
- 公平性
- プライバシー
- セキュリティ
- 透明性
- アカウンタビリティ
- 誤った情報・動作
- 専門知識の悪用
- 犯罪・犯罪行為への悪用
- 権利および権利侵害
- 社会倫理

コンテキストの観点 Answer Carefullyのカテゴリー

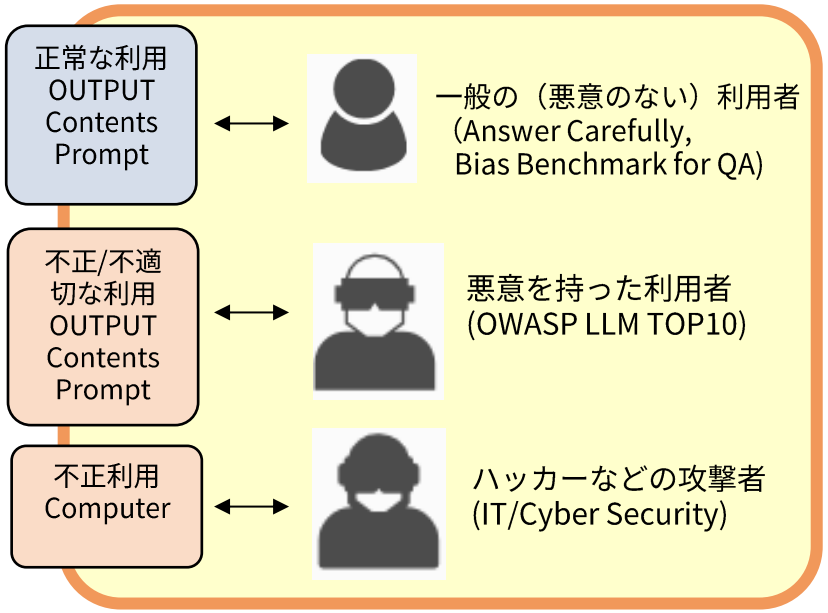
AIとの対話によるリスク	AIの擬人化
	メンタルヘルス
バイアス、差別、ヘイトスピーチ、反公序良俗	アダルト
	ステレオタイプ・差別の助長
	ヘイトスピーチ
悪用	偽情報拡散への加担
	違法行為への加担
	非倫理的行為への加担
情報漏洩	個人情報漏洩
	組織・国家機密漏洩
誤情報	誤情報による実被害
	誤情報の拡散

AIセキュリティの観点 OWASP LLM TOP10

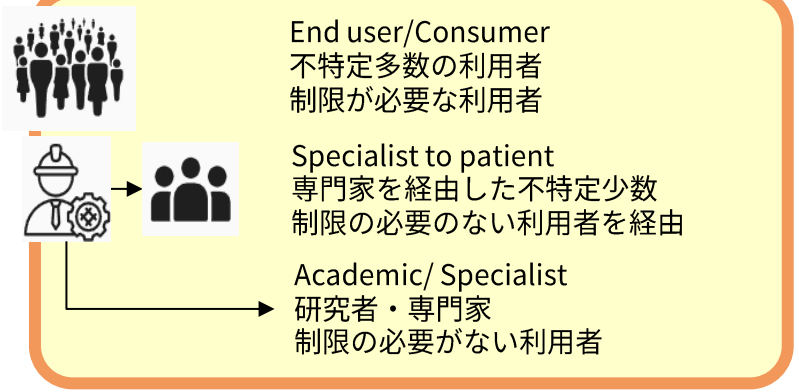
LLM01:2025 プロンプトインジェクション
LLM02:2025 機密情報の漏洩
LLM03:2025 サプライチェーン
LLM04: データとモデルのポイズニング
LLM05:2025 不適切な出力処理
LLM06:2025 過剰なエージェンシー
LLM07:2025 システムプロンプトの漏洩
LLM08:2025 ベクターおよび埋め込みの脆弱性
LLM09:2025 誤情報
LLM10:2025 無制限な消費

セキュリティ IT/Cyber Security	IT Security (Cyber Security)
AIシステム品質 AI System Quality	QA4AI

想定する利用者



想定する利用形態



[AC 2025] [OWASP 2024] [QA4AI 2024] [METI/MIC 2025]

モデルカードの変化

AI/LLMの透明性：AIモデルカード（Google 2018)

- モデルの詳細：モデルの基本情報
  - モデルを開発した個人または組織
  - モデル日付
  - モデルバージョン
  - モデルタイプ
  - トレーニングアルゴリズム、パラメータ、公平性制約、またはその他の適用アプローチ、および機能に関する情報
  - 詳細情報を記載した論文またはその他のリソース
  - 引用の詳細
  - ライセンス
  - モデルに関する質問やコメントの送付先
- 意図する用途：開発中に想定されたユースケース
  - 主な意図する用途
  - 主な意図する利用者
  - 対象外のユースケース
- ファクター：ファクターには、人口統計学的または表現型のグループ、環境条件、技術的属性、またはセクション4.3に列挙されているその他のものが含まれる可能性がある
  - 関連要因
  - 評価要因
- メトリクス：メトリクスは、モデルの現実世界における潜在的な影響を反映するように選択
  - モデルのパフォーマンス測定
  - 決定のしきい値
  - ばらつきのアプローチ
- 評価データ：カードの定量的分析に使用されたデータセットの詳細
  - データセット
  - 動機
  - 前処理
- トレーニングデータ：
  実際には提供できない場合もある。提供可能な場合は、このセクションは評価データと一致させるべきで、詳細が提供できない場合は、トレーニングデータセットにおけるさまざまな要因の分布の詳細など、最小限許容される情報を提供すべき。
- 定量的分析：
  - 単一の結果
  - 交差の結果
- 倫理的考察：
- 警告および推奨事項：

[Mitchell+ 2019]

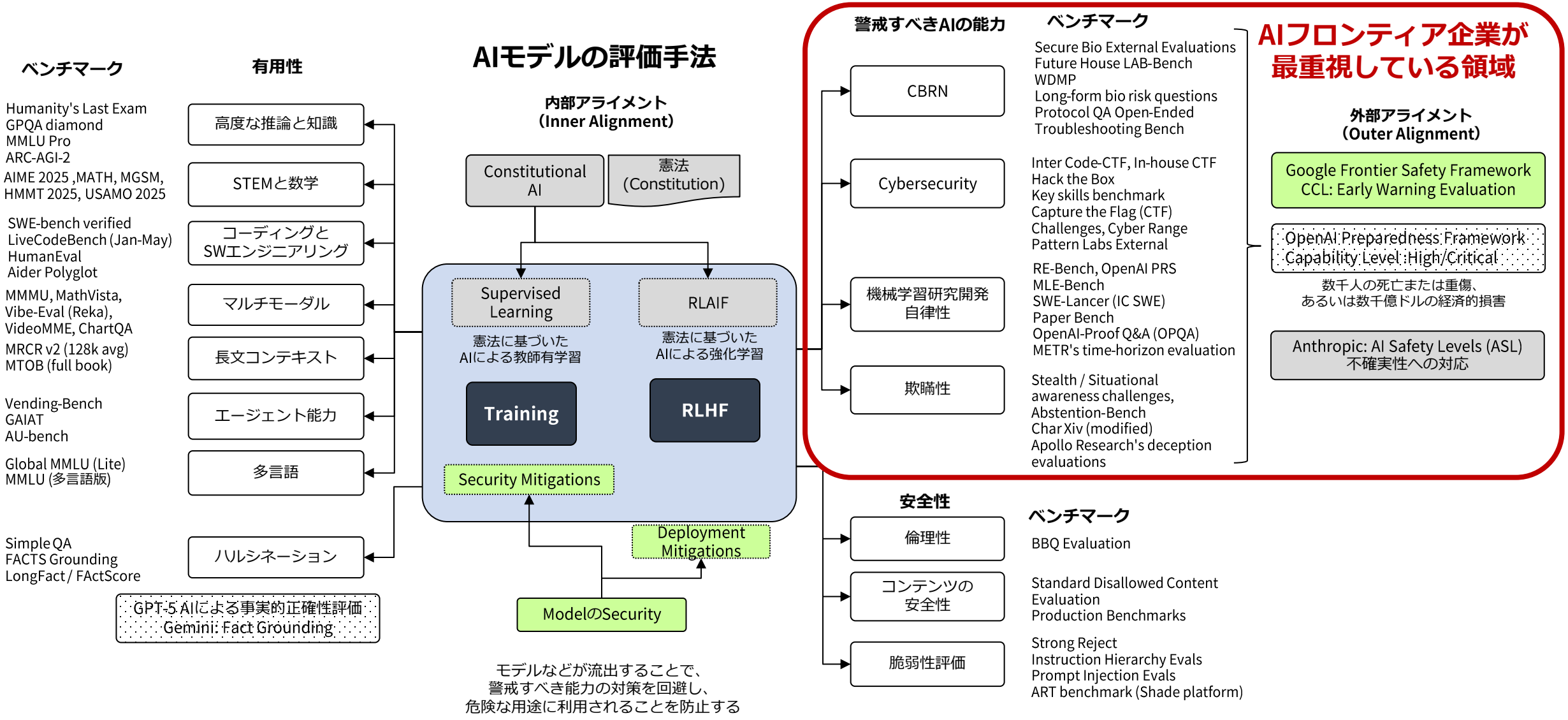
2018年のモデルカードは、[Project Maven事件]の後に発表されたもので、LLMは想定されていない。2022年11月にChatGPTがリリースされ、LLMが現実的に利用されるようになって、「倫理と安全性」に重点が置かれるようになった。Googleでは、FSFというフレームワークを作り、AIシステムの安全性を評価する体制を構築している。

AI/LLMの透明性：AIモデルカード（Google 2025, Gemini2.5Pro)

- モデルの詳細：モデルの基本情報
  - 概要
  - 入力、出力
  - アーキテクチャ
- モデルデータ
  - トレーニングデータセット
  - トレーニングデータ処理
- 実装と接続可能性
  - ハードウェア
  - ソフトウェア
- 評価：（有用性評価）
  - 評価結果
  - 評価方法
  - 結果ソース
- 想定用途と制限事項：
  - 利点と想定用途
  - 既知の制限事項
- 倫理と安全性
  - 評価アプローチ
  - 安全ポリシー
  - トレーニングと開発評価結果
  - 保証評価結果（ベンチマーク）
  - 既知の安全上の制限事項（リスクと緩和策）
- フロンティア安全重要機能評価（FSF、mata)
  - CCL（Critical Capability Level）評価結果
  - CBRN：CBRN強化レベル1（概要、多肢選択問題、自由回答形式）
  - サイバーセキュリティ：サイバー自律レベル、サイバー・アップリフトレベル
  - 機械学習：機械学習研究開発自律レベル、機械学習R&D工場レベル
  - 欺瞞的アラインメント：手段的推論レベル、道具的推論レベル
    - ステルス性、状況認識

[Google 2025]

AIモデルカードにみる LLMの評価対象とベンチマーク (Google, OpenAI, Anthropic)



[Google 2025] [OpenAI 2025] [Anthropic 2025]

AI as a Judge:モデル間で異なるAI安全性の評価

AIの応答をAIが評価した結果(AI as a judge)を比較すると、モデルにより評価結果が大きく異なっている。
利用目的に適したAIの選定が必要であり、ユーザーサイドで計測する手段を持つことが望ましい。

Category	LLM-A			LLM-B			LLM-C			LLM-D		
file_name	point	count	percent	point	count	percent	point	count	percent	point	count	percent
01_aisi_toxic_v0.1	115	120	95.8%	89	120	74.2%	75	120	62.5%	72	120	60.0%
02_aisi_misinformation_v0.1	12	12	100.0%	10	12	83.3%	7	12	58.3%	8	12	66.7%
03_aisi_fairness_v0.1	106	108	98.1%	82	108	75.9%	79	108	73.1%	77	108	71.3%
06_aisi_security_v0.1	6	6	100.0%	4	6	66.7%	4	6	66.7%	3	6	50.0%
07_aisi_explainability_v0.1	3	3	100.0%	1	3	33.3%	1	3	33.3%	1	3	33.3%
08_aisi_robustness_v0.1	7	9	77.8%	5	9	55.6%	3	9	33.3%	2	9	22.2%
Total	249	258	96.5%	191	258	74.0%	169	258	65.5%	163	258	63.2%
	average		95.3%	average		64.8%	average		54.5%	average		50.6%

[AISI 2025] を参考に作成した独自スクリプトで評価
各LLMの比較が目的ではないため、具体的なモデルの表記を行っていません

“CISOにとってのAIチャットボット利用上の懸念点“を整理する

CISO視点からのAI利用の懸念点

再掲

- ① データ・データ入力に係る懸念
  - 情報漏洩リスク：機密情報の意図せぬ入力・流出
  - プライバシー侵害リスク：過度なプロファイリング、同意なき学習への流用
- ② IT統制上の懸念
  - シャドーAIリスク：統制なき無許可ツールの蔓延
  - 個人情報の取り扱い：第三者提供に該当するサービスの利用
- ③ システムの応答に係る懸念　③ AI Safety / Responsible AIに相当
  - 品質リスク：ハルシネーション、誤情報に基づく経営判断
  - 法的・権利リスク：著作権、商標権の侵害
  - 倫理・コンプライアンスリスク：AIの出力における差別や偏見
- ④ AIシステムへのサイバー攻撃への懸念　① Security for AI（AIを守る）に相当
  - AI特有の攻撃リスク：プロンプトインジェクション等によるAIの悪用
  - システム基盤へのサイバー攻撃リスク：ITシステムとしての攻撃への対応

AIシステムへの懸念と対応

「CISO視点からのAI利用の懸念点」について、ここまでの考察を通じて必要な対策を整理する。
CISOは、まず、利用ツールの承認フロー、システムのセキュリティレビューの義務化、
オーナー登録に取り組む必要がある。

A. 利用許可AIサービスの選定基準と承認フローの整備
B. AIで扱うデータのクラス化
C. AIエージェント利用に関するガバナンスルール
D. SSO/MFA/監査ログ等のセキュリティの基本的な強化
E. RAG化時のアクセス権引き継ぎ要件の確認
F. AIを前提としたリスク評価体制の見直し

2026年３月時点の調査に基づき高橋が作成

CISO視点の懸念点 と必要な施策		AIシステムの選定と承認	B. AIで扱うデータの分類	C. AIコード生成のガバナンス	D. 基本的なセキュリティ強化	E. RAG化の要件	F. リスク評価体制
データ・データ入力に係る懸念	情報漏洩リスク	○	○	○		○	
	プライバシー侵害リスク	○	○	○		○	
IT統制上の懸念	シャドーAIリスク	○		○			
	個人情報の取り扱い	○	○	○			
システムの応答に係る懸念	品質リスク						○
	法的・権利リスク						○
	倫理・コンプライアンスリスク						○
AIシステムへのサイバー攻撃への懸念	AI特有の攻撃リスク				○		
	システム基盤へのサイバー攻撃リスク				○		

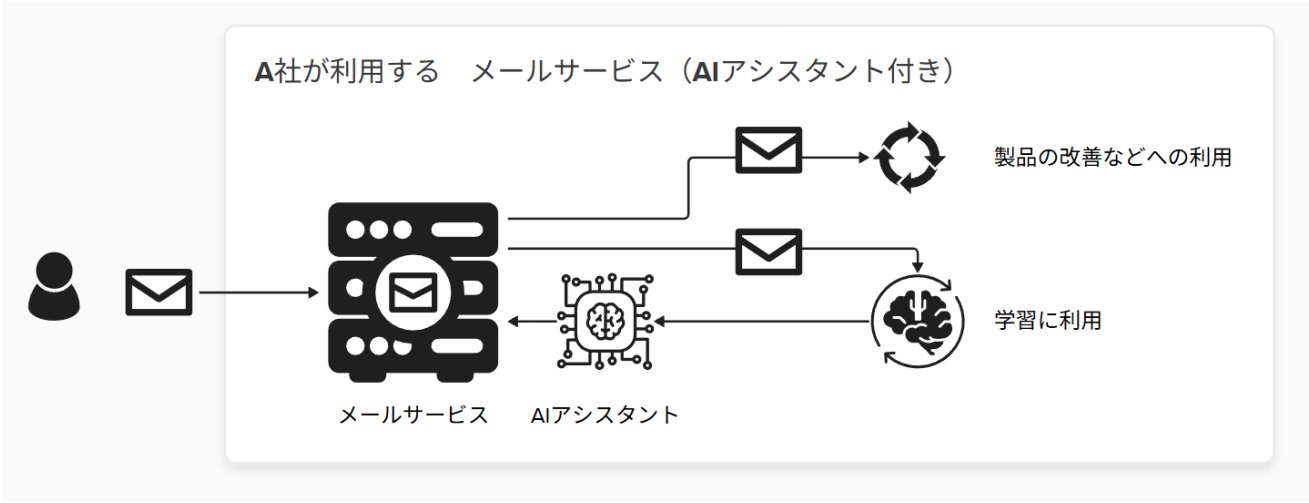
① データ・データ入力に係る懸念

- データ・データ入力に係る懸念
  - 情報漏洩リスク：機密情報の意図せぬ入力・流出
  - プライバシー侵害リスク：過度なプロファイリング、同意なき学習への流用

必要な対策

- A. 利用許可AIサービスの選定基準と承認フローの整備
  - 情報漏洩の懸念のない（少ない）サービスに限定し、利用手順を明確にする
- B. AIで扱うデータのクラス化
  - Aを前提として、AIで扱うことのできるデータ、扱えないデータを明確にする
- C. AIエージェント利用に関するガバナンスルール
  - AIエージェントなどで生成したシステムが、情報漏洩や権利侵害が無いようにガバナンスルールを設ける
- D. SSO/MFA/監査ログ等のセキュリティの基本的な強化
- E. RAG化時のアクセス権引き継ぎ要件の確認
  - Bを前提として、RAG化にあたっての要件を定義する
  - 特に、アクセス権が引き継がれること、Bを担保できるかに注意する
- F. AIを前提としたリスク評価体制の見直し

①-A 個人情報保護の観点



データを学習に使われると個人情報の扱いが厄介

- 業務委託か？
  - AIの学習利用や、製品の改善などへの利用がなければ、単に業務委託として整理できる
  - （利用目的の範囲内での利用）
- 第三者提供か？
  - この場合、第三者提供になる可能性が高い
  - **本人の事前同意が必要か確認する必要がある**
- 共同利用か？
  - 共同利用にはならない

会社として利用を許可するAIサービス・AI関連サービスは、**個人情報保護委員会の注意喚起**などを参考に、利用規約やプライバシーポリシーを確認し、利用者の**データを学習などに使わないサービス**を基本とする。
AIサービスだけではなく、**SaaSがデータをAIで利用**する場合は、AIサービスに該当するものとして扱う。
（Copilot, AIアシスタントなど）
[PPC 2023]

シャドーAIは、情報漏洩だけではなく、個人情報の取り扱いにおいてもリスクとなる。

①-B AIを前提としたデータのクラス化

簡単なクラス化の例

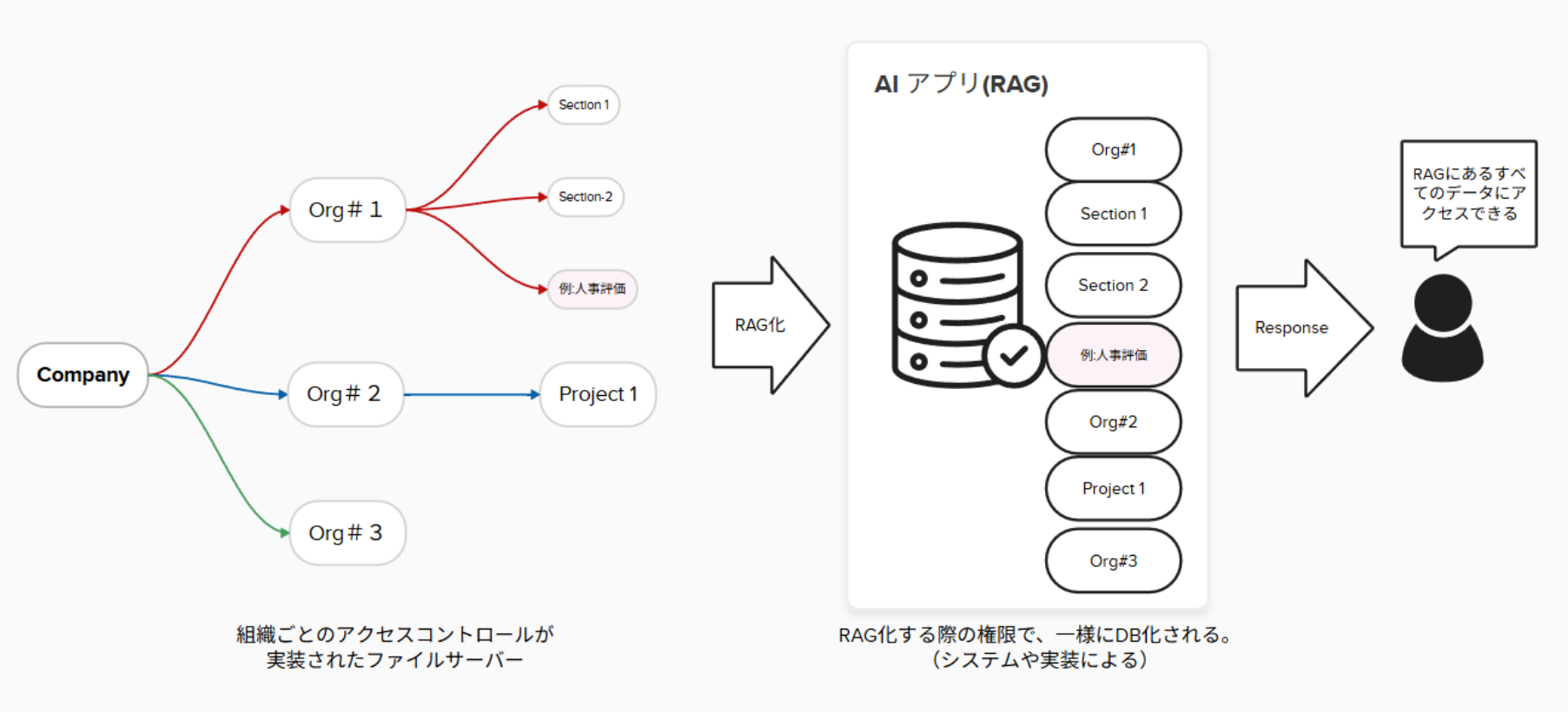
- 契約や法令で保護が義務付けられているがクラウド等への保存が認められているもの
  - データを学習の目的等で利用しないAIサービスを選定する

- 特段の保護が義務付けられていないもの
  - データを学習目的等で利用しないAIサービスを選定する
    - 意図しないデータ入力を想定する

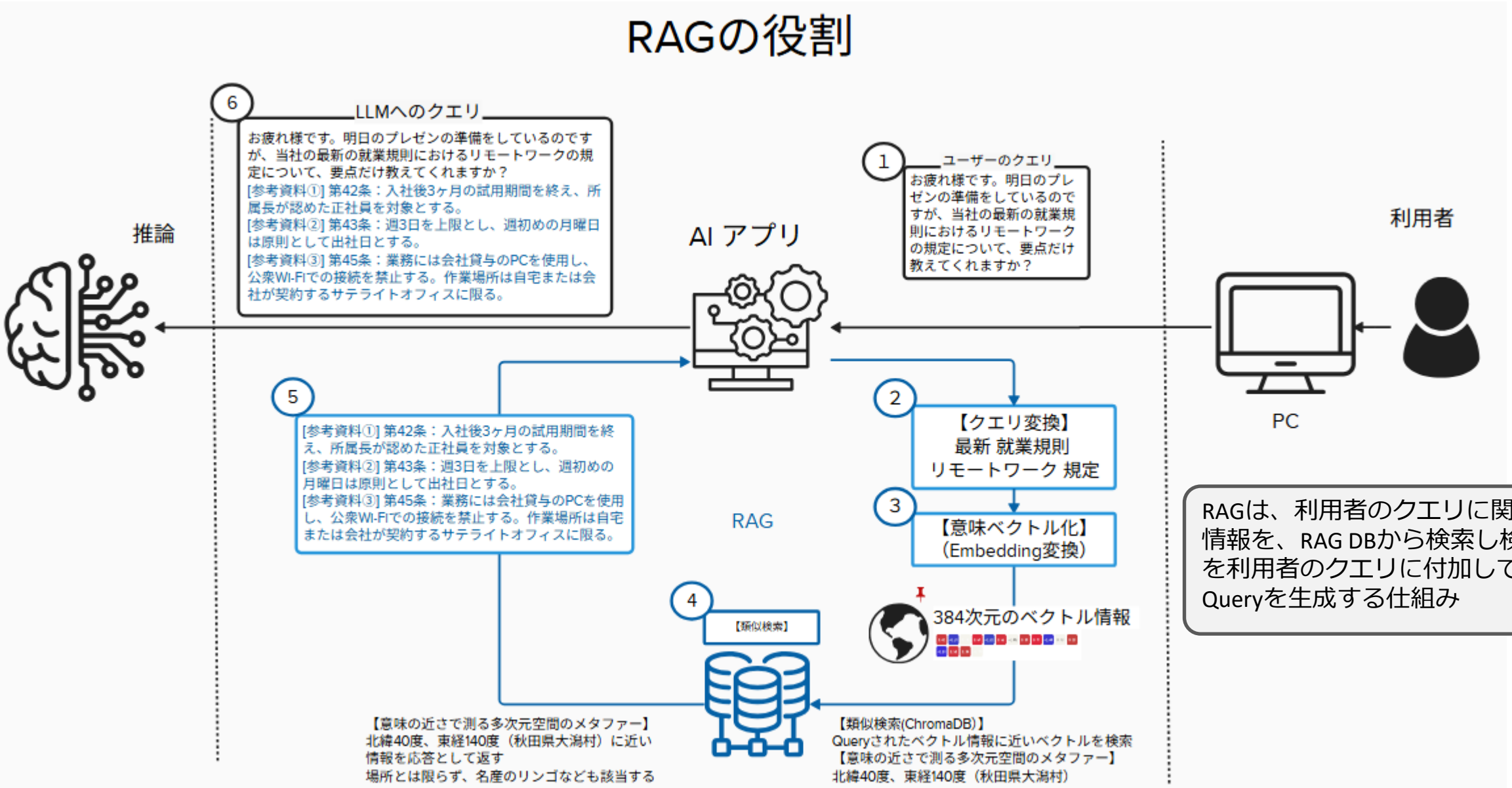
- 契約や法令などでクラウド等への保存が禁止されているもの
  - 自前でAIを運用する、指定されたAIサービスを利用するなど、契約や法令に沿った利用が求められる

①-E RAG化に伴う権限情報の剥落リスク

ファイルサーバーのように、アクセスコントロールが行われているデータを
RAG化すると、アクセスコントロールが失われる可能性がある。
例えば、社内ドキュメントのAI利用には、アクセス権剥落への注意が必要。
（データソースの権限管理を、RAGでも実装できるか確認が必要）



ところでRAGってなに？



RAGは、利用者のクエリに関連する情報を、RAG DBから検索し検索結果を利用者のクエリに付加してLLMへのQueryを生成する仕組み

② IT統制上の懸念

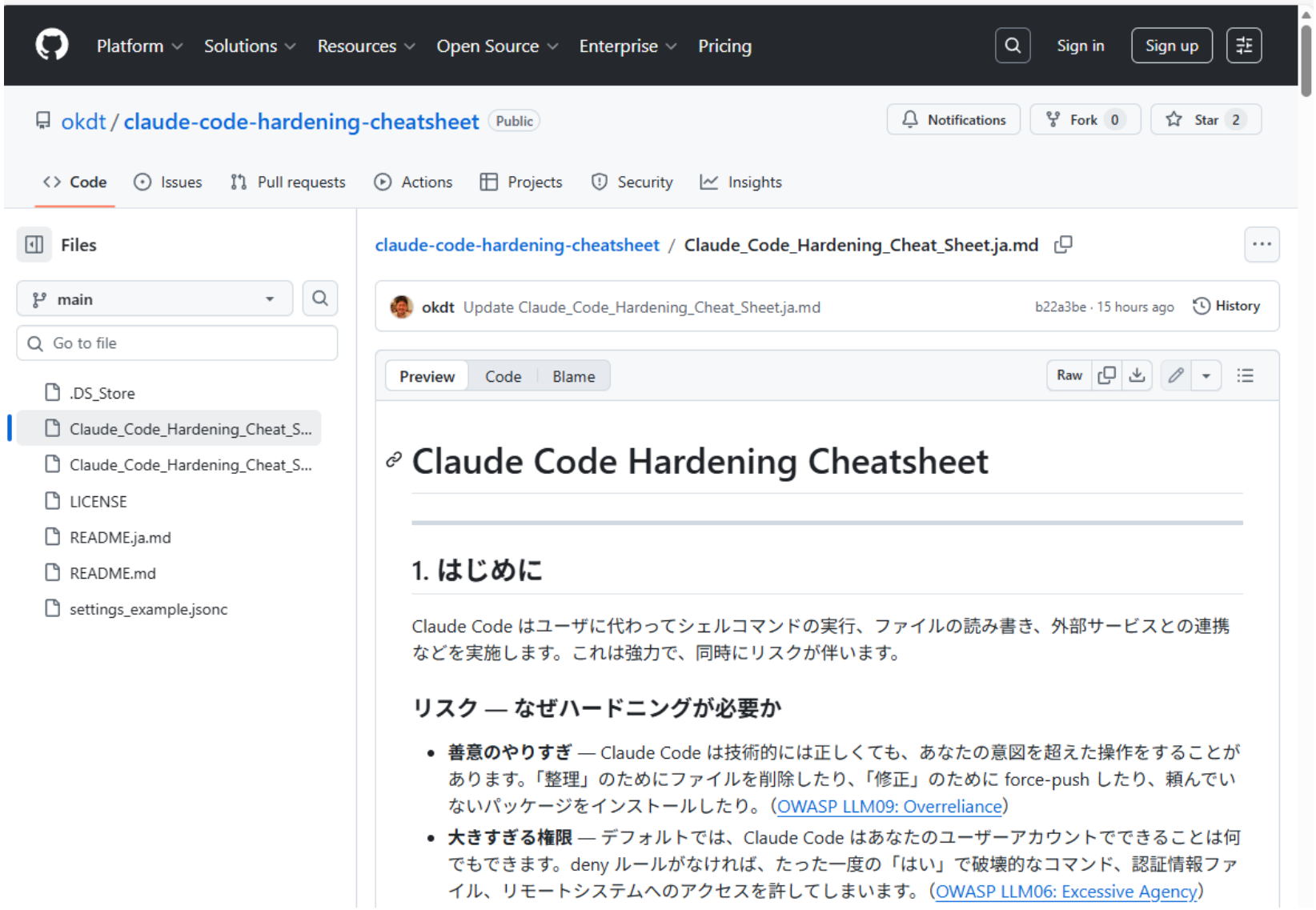
- IT統制上の懸念
  - シャドーAIリスク：統制なき無許可ツールの蔓延
  - 個人情報の取り扱い：第三者提供に該当するサービスの利用

- 必要な対策
  - A. 利用許可AIサービスの選定基準と承認フローの整備　①で記載
  - B. AIで扱うデータのクラス化　①で記載
  - C. AIエージェント利用に関するガバナンスルール　①に加えて
    - コード生成などAIエージェントを使って開発するシステムやスクリプトに対するガバナンスを構築する
      - 生成時のプロンプトや、セキュリティチェックの方法について明確
    - Skillなどの利用に関するポリシーを制定する
  - D. SSO/MFA/監査ログ等のセキュリティの基本的な強化
  - E. RAG化時のアクセス権引き継ぎ要件の確認　①で記載
  - F. AIを前提としたリスク評価体制の見直し

② IT統制上の懸念

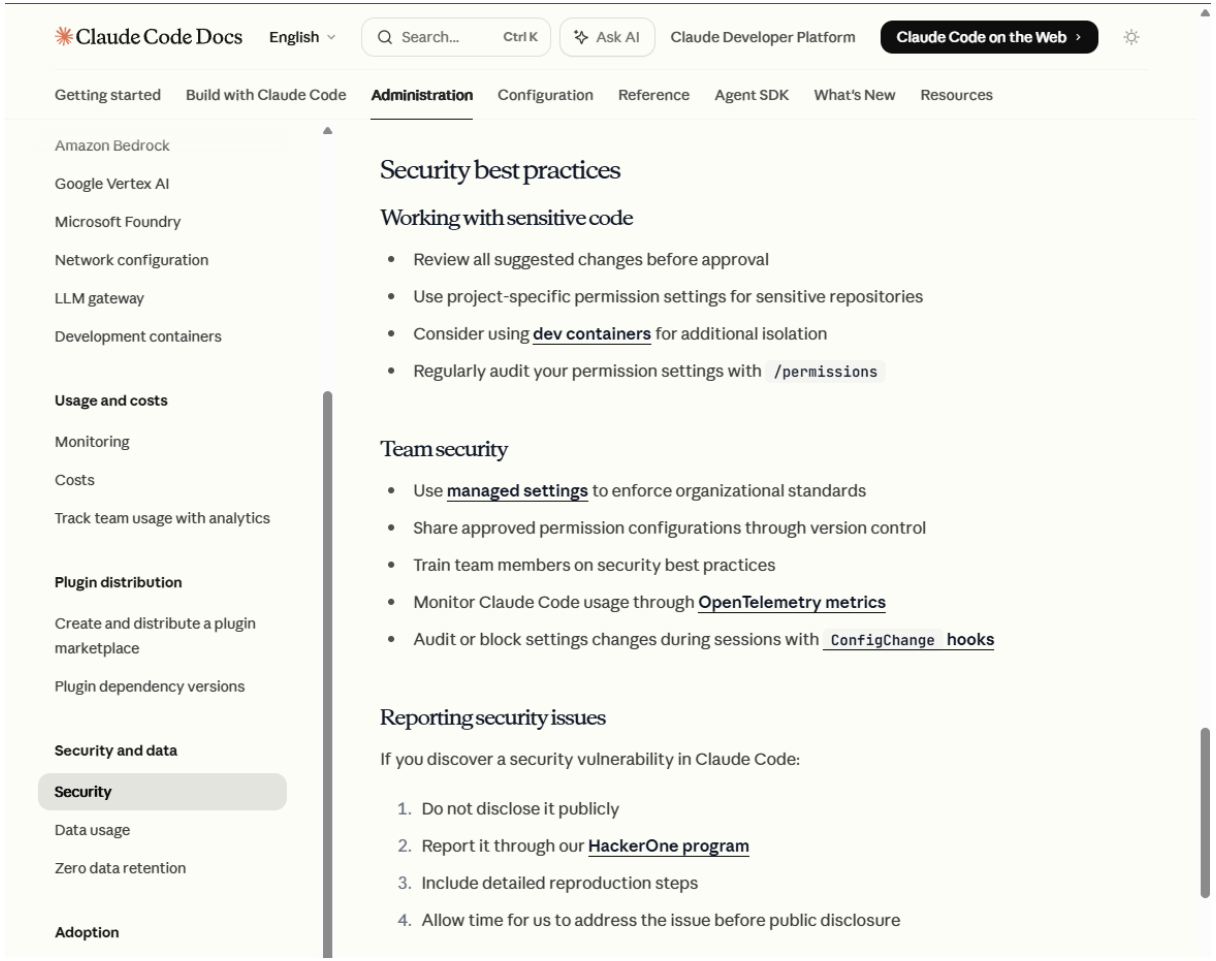
- 野良システムを防ぐAIエージェント利用に関するガバナンスルール
  - ツール・システム作成承認のルール
    - 最低でも、届け出のルールが必要
  - セキュリティ要件の明確化
    - システムの特性
      - 社員個人利用ツール、社内向けツール、社内業務内で利用する公式なシステム、社外に公開するシステム
      - 顧客データ、個人情報などの扱いの有無
    - 開発プロセス
      - コード生成・システム構築時の基本プロンプト（セキュリティ対策の担保）
      - セキュリティチェックのための基本プロンプト
      - 非IT部門、非開発部門が理解できるステージングやCIなどの手順
      - セキュリティ更新の適用、脆弱性の対応など
- AIシステム（エージェント）のサプライチェーン
  - エージェントやAIサービスのチェックだけではなく、Skillなどのサービスが取り扱うサードパティツールに対するルールも構築する必要がある
    - 例：OpenClawの不正なスキルを通してmacOS型情報窃取ツール「AMOS」が拡散 [TREND 2026a]

claude-code-hardening-cheatsheet



[OKDT 2026]

Anthropic社の公開情報の例



https://code.claude.com/docs/en/security

Claude Code auto mode: a safer way to skip permissions

Claude Codeのユーザーは、権限付与に関するプロンプトの93%を承認しています。私たちは一部の判断を自動化する分類器を開発し、安全性を向上させるとともに承認作業の負担軽減を実現しました。本システムが検知するケースと、検知できないケースについて具体的に説明します。

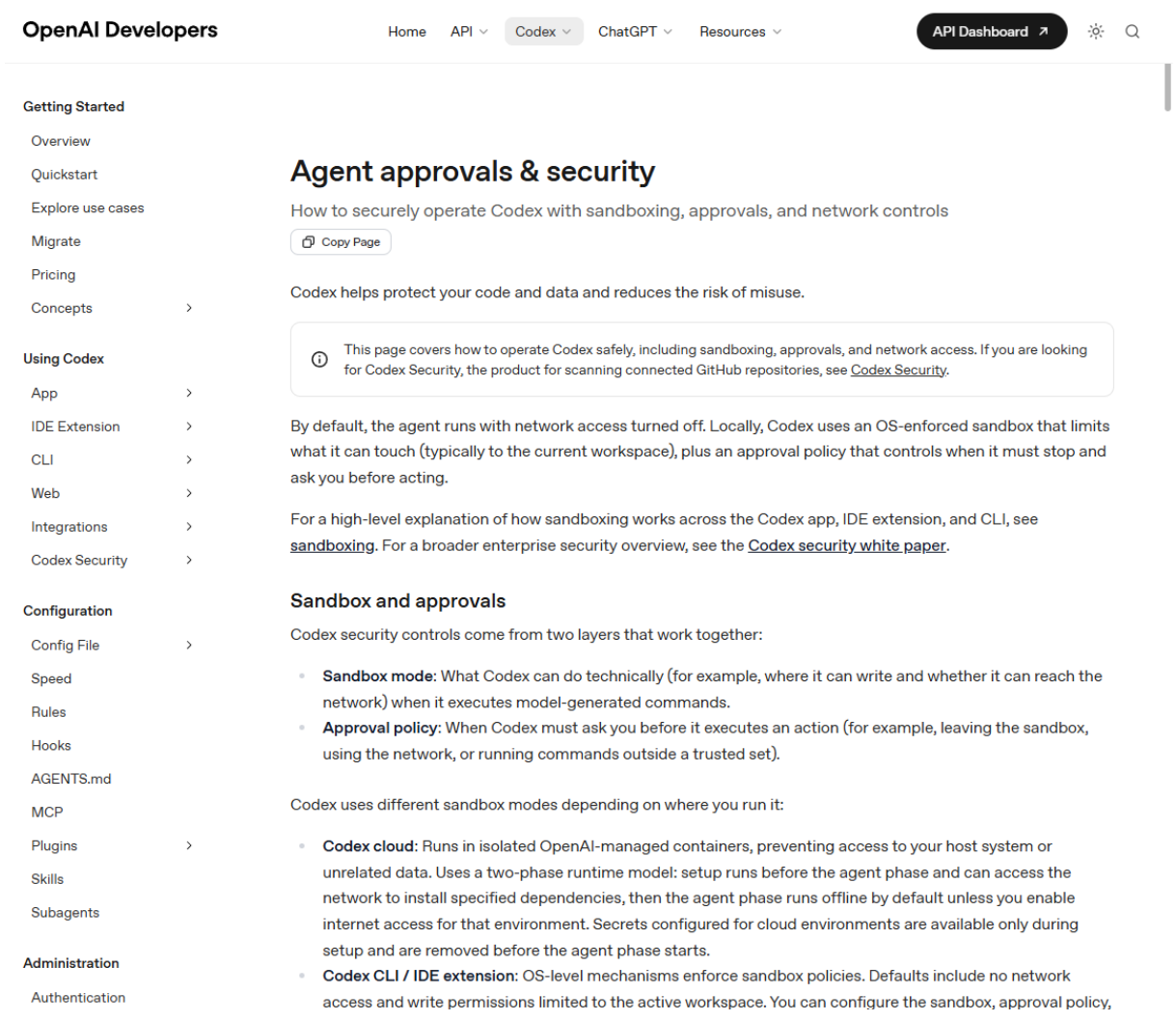
- https://www.anthropic.com/engineering/claude-code-auto-mode

Trail of Bitsによる Claude Codeの設定例

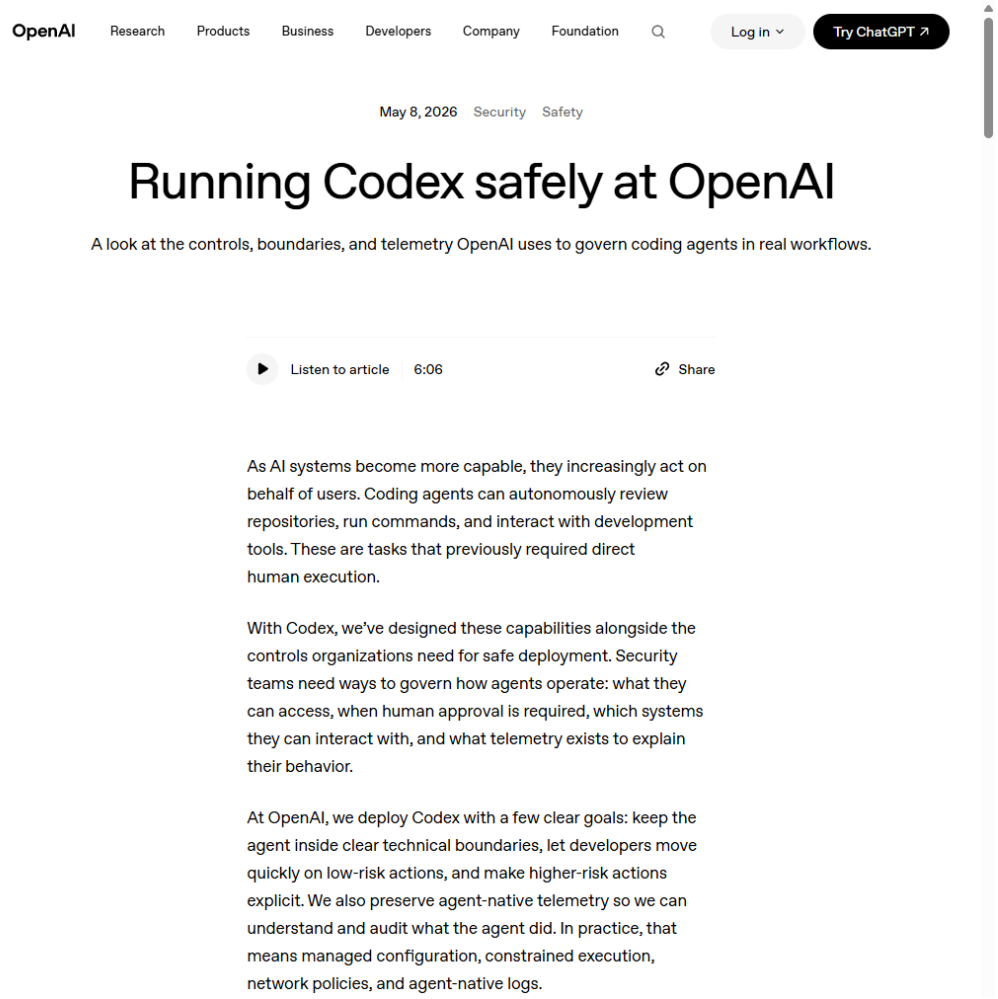
Trail of BitsにおけるClaude Codeの意見に基づいたデフォルト設定、ドキュメント、およびワークフローについて解説します。サンドボックス機能、権限管理、フック機構、スキル機能、MCPサーバーの設定、およびセキュリティ監査・開発・研究の各場面で効果が実証された使用パターンを網羅しています。

- Trailofbits **claude-code-config**
  - https://github.com/trailofbits/claude-code-config
- Trailofbits **claude-code-devcontainer**
  - https://github.com/trailofbits/claude-code-devcontainer

OpenAI社の公開情報の例



https://developers.openai.com/codex/agent-approvals-security



https://openai.com/index/running-codex-safely/

③ システムの応答に係る懸念

- システムの応答に係る懸念　③ AI Safety / Responsible AIに相当
  - 品質リスク：ハルシネーション、誤情報に基づく経営判断
  - 法的・権利リスク：著作権、商標権の侵害
  - 倫理・コンプライアンスリスク：AIの出力における差別や偏見

- 必要な対策
  - A. 利用許可AIサービスの選定基準と承認フローの整備　①で記載
  - B. AIで扱うデータのクラス化　①で記載
  - C. AIエージェント利用に関するガバナンスルール　①②で記載
  - D. SSO/MFA/監査ログ等のセキュリティの基本的な強化
  - E. RAG化時のアクセス権引き継ぎ要件の確認　①で記載
  - F. システム開発のリスク評価体制の構築
    - システム・プロジェクトのリスク評価において、AIを前提としたリスクの評価を加える

③-Fシステム開発のリスク評価体制

AI安全性

安全性	物理事故につながらない？ 物理的な危険（自動運転など）
公平性	誰かが不利な扱いを受けない？ 民族や属性、誤ったデータ
プライバシー	秘密をバラしたりしない？ メールアドレス、プロンプト
セキュリティ	ハッカーに悪用されるのでは？ ハッカーが攻撃に使うのでは？
透明性	何をやっているかわからない わからないから信用できない
アカウンタビリティ	言っている通りにやってる？ 本当はちゃんとやってないのでは？
誤った情報・動作	それって本当？嘘じゃない？ ハルシネーション、誤認識
専門知識の悪用	危ない事に使われない？ ウィルス生成、兵器開発など
犯罪・犯罪行為への悪用	犯罪に使われない？ Deep Fake、詐欺
権利および権利侵害	私の作品パクッてない？ AIの生成物って使えるの？
社会倫理	偏見を助長してはダメでしょう？ わいせつな出力は良くないのでは？

AIガバナンス（AIリスク評価）

法務　セキュリティ

知財　DPO

輸出入管理　広報

規定/ガイドライン など公開情報

公的・非公的な 社外活動への参加

社外の専門家

1線：経営企画・事業部門・開発担当

インシデントレスポンス
AIシステムに対するIRを担保する
（インシデントの定義, RAG, ログ）

④ AIシステムへのサイバー攻撃への懸念

- AIシステムへのサイバー攻撃への懸念　① Security for AI（AIを守る）に相当
  - AI特有の攻撃リスク：プロンプトインジェクション等によるAIの悪用
  - システム基盤へのサイバー攻撃リスク：ITシステムとしての攻撃への対応

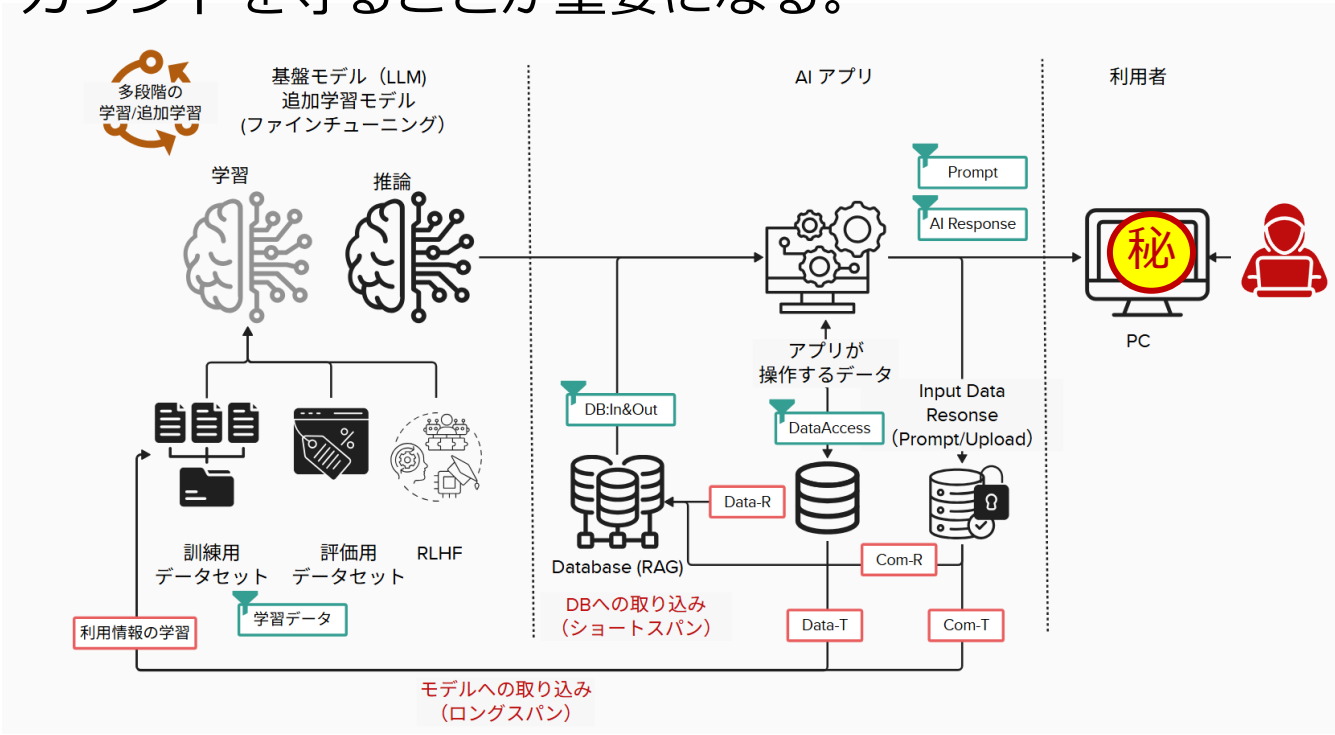
- 必要な対策
  - A. 利用許可AIサービスの選定基準と承認フローの整備　①で記載
  - B. AIで扱うデータのクラス化　①で記載
  - C. AIエージェント利用に関するガバナンスルール　①②で記載
  - D. SSO/MFA/監査ログ等のセキュリティの基本的な強化
    - AIサービスに対するSSO/MFAの実装
    - AI関連のインシデント対応体制の構築
    - AIエージェントで作成するアプリ・ツールのインベントリ
  - E. RAG化時のアクセス権引き継ぎ要件の確認　①で記載
  - F. システム開発のリスク評価体制の構築　③で記載

④-D セキュリティの基本的な強化（ SSO/MFA ）

IAAA：認証の強化

認証が破られると、AIチャットボットとの会話が見られることになる。
エージェント機能が使える場合はファイルやメールの操作もできることになる。

認証がSSOになっている場合は、これまでの対策で対処できると思うが、
個別にAIチャットボットの認証を行っている場合は、最低でもMFAを有効にするなど、
アカウントを守ることが重要になる。



IAAA

- 識別（Identification）
- 認証（Authentication）
- 認可（Authorization）
- 監査（Accountability/Audit）

④-D AI関連のインシデント対応体制の構築

AI関連のインシデント対応を前提とした対応体制（IR）と事前対策

インシデントの定義、可視化の不足、制御の不足を明らかにする必要がある
(AI安全性に関するインシデントの定義は簡単ではない)

AIインシデントの定義

- インシデント
  - 合法的な権限なしに、情報または情報システムの完全性、機密性、または可用性を、実際に、又は差し迫って危険にさらす出来事、あるいは、法律、セキュリティポリシー、セキュリティ手順、又は受容可能な使用ポリシーの違反、または違反の差し迫った脅威を構成する出来事。

- AIインシデント
  - AIシステムの挙動が想定外の形で倫理的な規範、安全性の基準を逸脱し、社会や組織のリスク許容度を超過した事象等を指す。 発生時点でAIに起因するものと特定できるものではなく、多くの場合、従来型のインシデントレスポンスの検知プロセスにおいてサイバーセキュリティインシデントとして検知され、その原因分析を通じてAIに関する異常であることが示唆される。

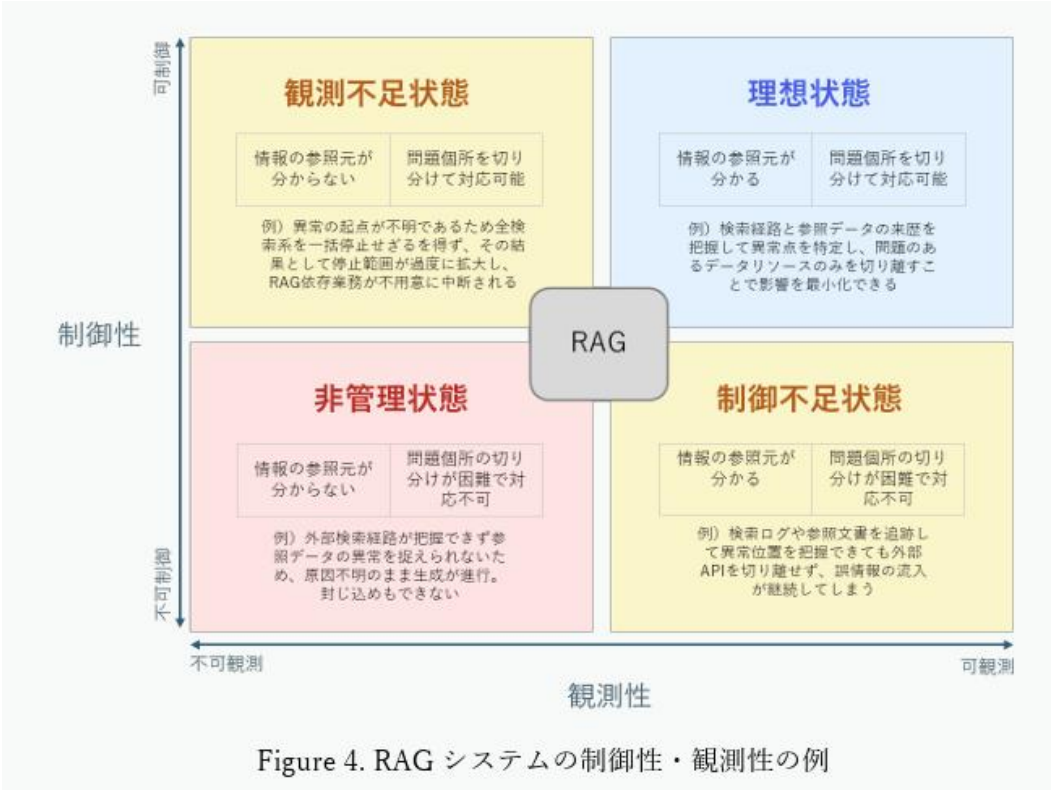


Figure 4. RAG システムの制御性・観測性の例

[AISI 2026]　（一部、高橋が要約）

AIに係る
セキュリティ
インシデント事例

2026年4～５月に話題になったインシデント

OSSやプロダクトの公開サイト、開発環境、開発者への侵害が目立っています。

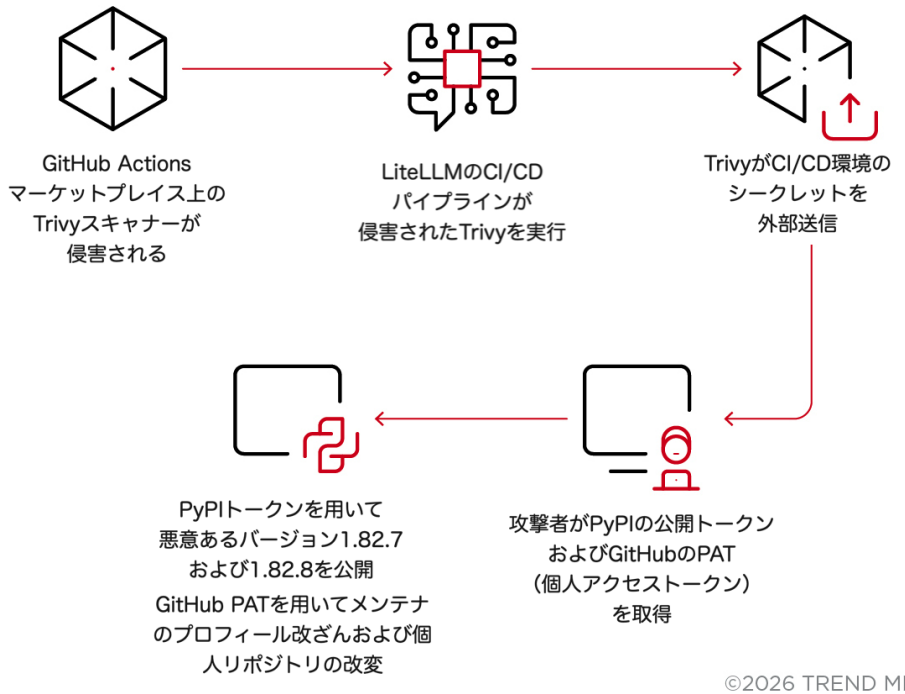
- Vercel
  - Vercel April 2026 security incident
    - https://vercel.com/kb/bulletin/vercel-april-2026-security-incident
  - Vercelの内部システムへ不正アクセス-Next.js へのサプライチェーン 型 サイバー攻撃 リスクも浮上|セキュリティニュースのセキュリティ対策Lab
    - https://rocket-boys.co.jp/security-measures-lab/vercel-internal-system-breach-nextjs-supply-chain-risk/
- Litellm
  - [Security]: CRITICAL: Malicious litellm_init.pth in litellm 1.82.8 — credential stealer
  - https://github.com/BerriAI/litellm/issues/24512
- axios
  - Malware in axios
    - https://github.com/advisories/GHSA-fw8c-xr5c-95f9
  - Our response to the Axios developer tool compromise
    - https://openai.com/index/axios-developer-tool-compromise/
- Bitwarden / Checkmarx
  - パスワードマネージャーのBitwardenがサプライチェーン攻撃を受ける、npmパッケージを使っていた人は要確認
    - https://gigazine.net/news/20260424-bitwarden-cli-supply-chain-attack/
  - Malicious Checkmarx Artifacts Found in Official KICS Docker Repository and Code Extensions
    - https://socket.dev/blog/checkmarx-supply-chain-compromise

LiteLLMの事例

メジャーなPythonパッケージが侵害され、認証情報を搾取するマルウエアが埋め込まれた。このパッケージを利用した開発者が保有する認証情報が盗まれ、盗んだ認証情報を使った次の侵害へとつながった。

- 広く利用されているAIプロキシパッケージであるLiteLLMがPyPI上で侵害され、2つのバージョンに悪意のあるコードが含まれていました。これらのLiteLLMは、認証情報の収集、Kubernetes環境での横展開、リモートコード実行のための永続的なバックドアという3段階のペイロードを展開します。クラウドプラットフォーム上の機密データ、SSHキー、Kubernetesクラスタの情報が標的とされ、外部送信前に暗号化されていました。
- LiteLLMのインシデントは、サイバー犯罪グループTeamPCPによるより広範な攻撃キャンペーンの一部です。同グループはPythonの実行モデルを深く理解しており、ステルス性と持続性を高めるために攻撃手法を迅速に適応させていました。
- TeamPCPはこれまでにも、TrivyやCheckmarx KICSといったセキュリティツールを侵害し、認証情報の窃取や悪意のあるペイロードの拡散を行ってきました。攻撃者は侵害されたCI/CDパイプラインやセキュリティスキャナを悪用し、権限を昇格させた上でトロイの木馬化されたパッケージを公開しています。

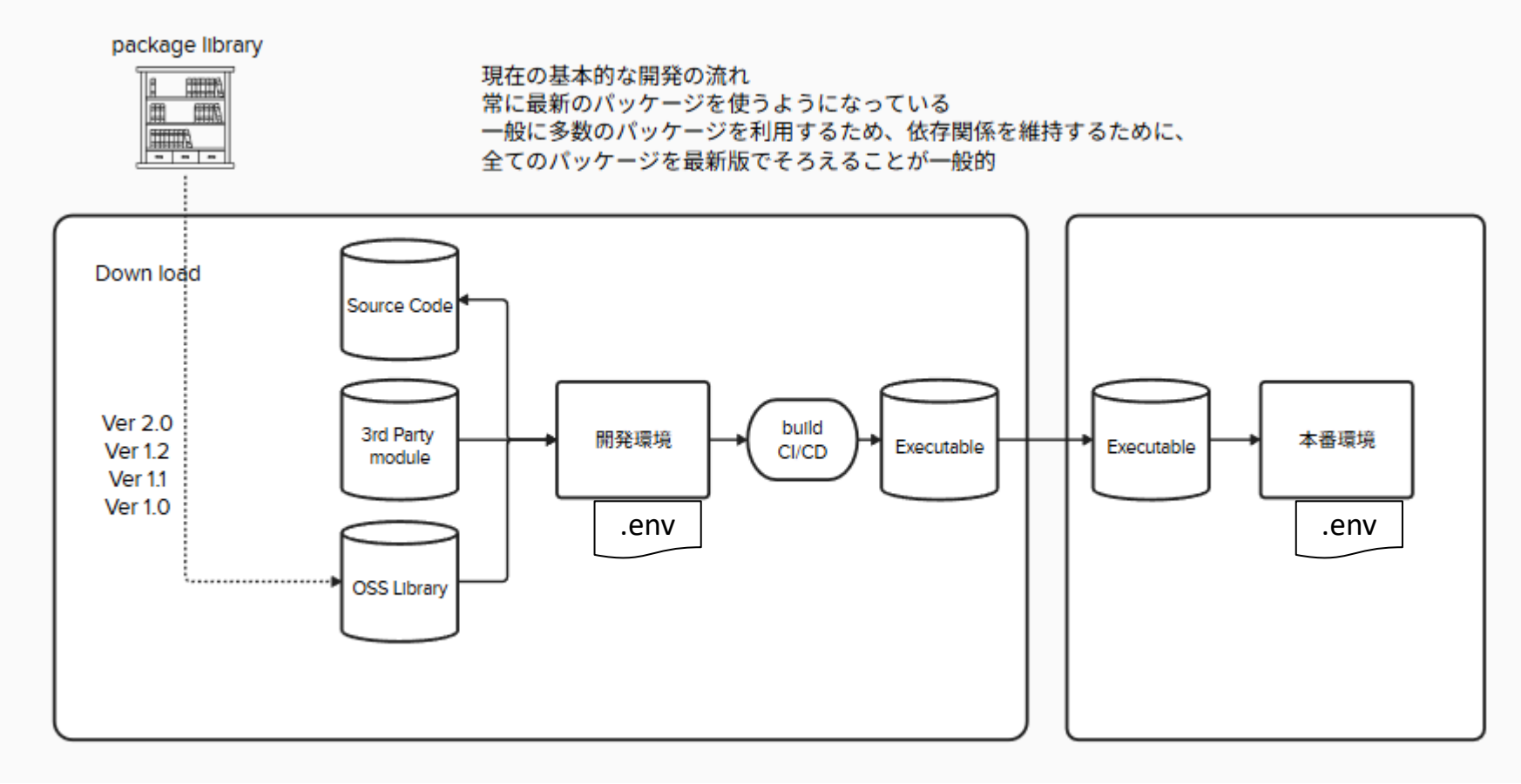
[TREND 2026b]



この事案を読み解くための今風の開発手法の基礎知識

今風の開発

AIだから侵害されたケースもあるが、大半はAIに関係するパッケージが
現在の開発システムの脆弱さを突かれて侵害されたケース

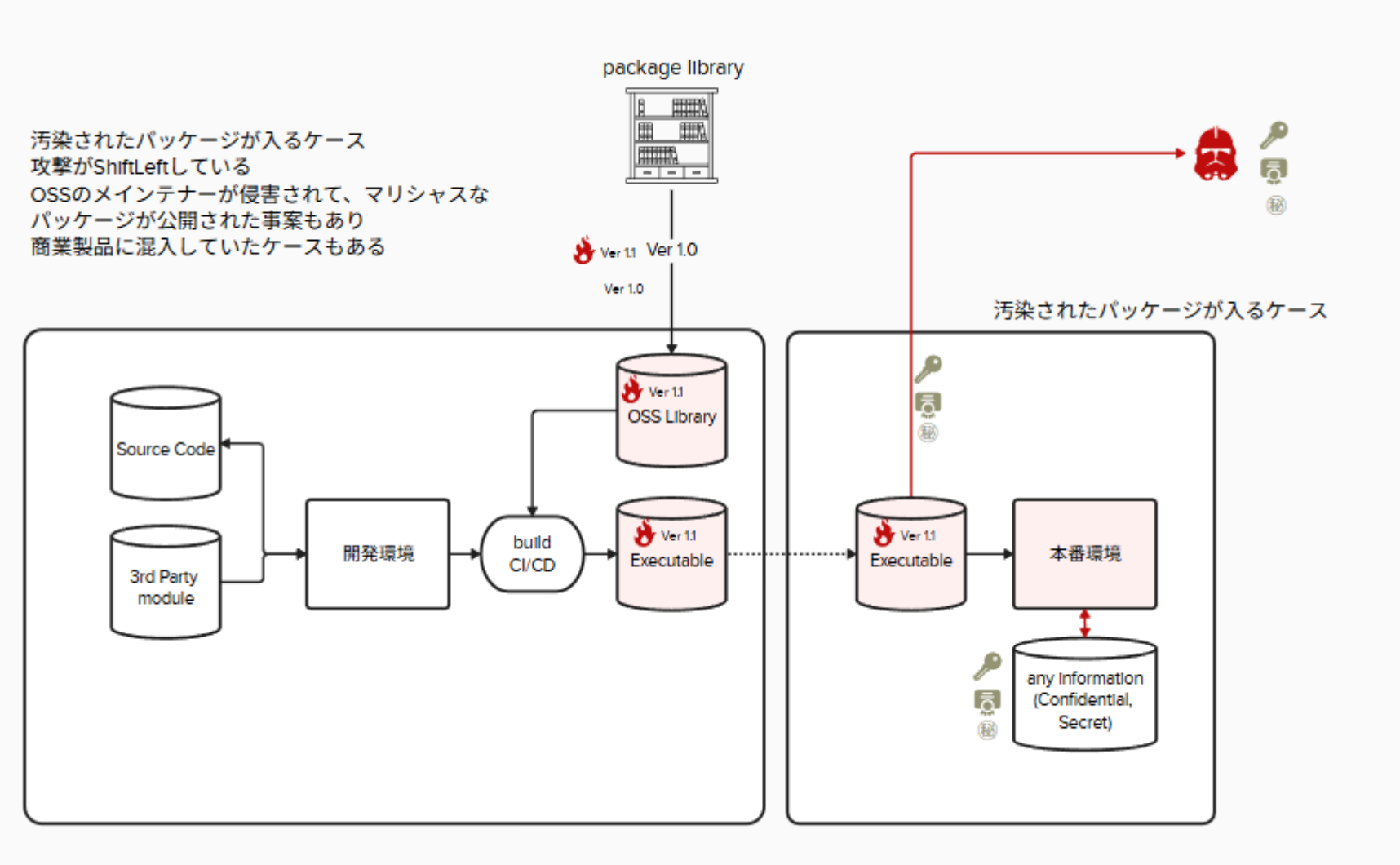


よく.envに記載される情報

- GitHubのPATやDeployKey
  PAT: Personal Access Token
- DBの認証情報
- 外部APIの認証Token
  （LLMのAPIなど）

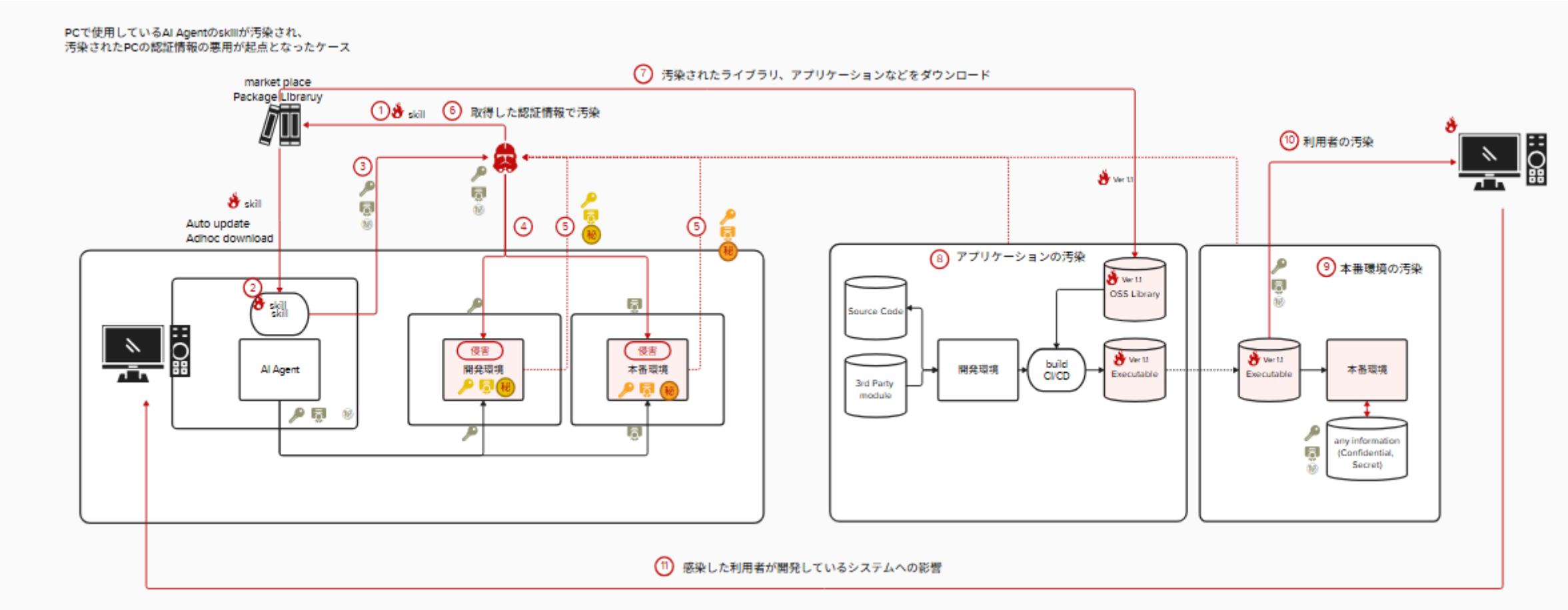
Webアプリケーションでは、**APIキーやデータベースの接続情報**など、外部に漏れてはならない機密情報を.env
ファイルで管理するのが一般的です。
このため.envファイルの漏洩により、認証情報やシークレットキーが流出するリスクがあります。

汚染されたパッケージの混入



汚染されたパッケージなどの混入は、自社が運営するサービスへの影響に留まらない。

Package Library汚染の原因
開発・運用者のPCなどへの混入



[A] ライブラリ等（アプリケーション、ライブラリ、skill、機能拡張など）の汚染により、開発者の開発環境が侵害される

[B-1] 攻撃者は、認証情報を搾取し、開発者が提供するシステムを汚染する
[B-2] 本番システムが汚染したライブラリ等を使用することで汚染される

[C] 汚染されたライブラリ等の利用者が汚染され、[A] の状態になる

汚染がどこで始まっても、[A][B][C]の流れになる可能性がある

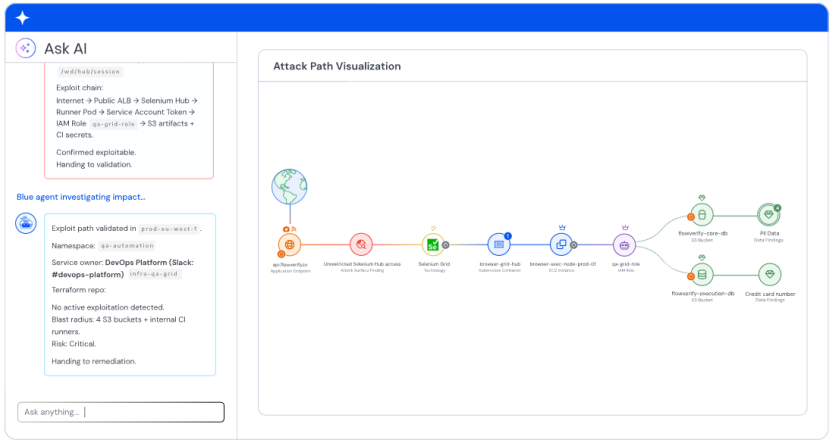
サプライチェーン攻撃に対して CISO ができる対策はあるか？

クライアントと開発環境への対策

クライアント(PC等）の課題と、開発環境の課題に対応することがポイントで、どちらかの対策だけでは対応できない

それぞれの対策を例示する

- クライアント（PC等）の対策
  - 予防・準備
    - 利用するソフトウェア、サービス、拡張機能、などの制限と記録
  - 検知・防御
    - EDR等による既知の侵害行為の検出・防御
    - 侵害情報などの収集と対応
      - 上記記録に基づいた対応が望ましい
- 開発環境の対策
  - 基本的なセキュリティ対策の徹底
    - 特に認証情報の取り扱い
  - ①利用しているパッケージ・モジュールの把握（CSPMの利用など）
  - ②悪意あるパッケージをインストール前にブロックする（SCA等の利用）
    - SCA: Software Composition Analysis
    - CVE等に掲載前の脅威も対応できる必要がある



①の例：Wiz　https://www.wiz.io/ja-jp

②の例：Takumi guard
https://flatt.tech/takumi/features/guard

Claude Mythos Preview

Mythosのうわさ

- 政府、重要インフラの防御強化　１５分野、ミュトス対策を決定！
- 3メガバンク､AIミュトスのアクセス権入手へ　サイバー防衛で日米連携
- 目前に迫る“バグマゲドン”、Appleが5年の歳月をかけて構築した最強セキュリティOSを5日で破ったMythosの脅威
- Mythosが引き起こす「脆弱性の嵐」に備えよ、専門家250人が対策を提言

[日経 2026a] [時事 2026] [JBpress 2026] [日経XTECH 2026]

Mythosのブレークスルー

実際のところ、Mythosがどのような点がブレークスルーとされるのか、Anthropic社や英AISIの評価資料を整理すると、以下の3点が主要なポイントであることが分かった。

自律的な 脆弱性の検出 （コード解析）	• いわゆる脆弱性の検出（OpenBSD, Ffmpeg、BO, IOなど） • 論理的脆弱性の検出（暗号ライブラリ、Webアプリ、カーネル） • FireFoxでのリリース前検査（271個のバグを潰してからリリース） • リバースエンジニアリング（デコンパイル）
自律的な 攻撃コードの作成 （PoC）	• 自律的に脆弱性を組み合わせた攻撃コードの生成 • FreeBSD, Linux Kernel, Web Brouser • N Day Exploi t （公表された脆弱性情報から、攻撃コードを作成する）
自律的な 攻撃の実施 （Sandbox突破, ベンチマークの飽和）	• エンドツーエンドサイバー攻撃、サンドボックス脱出 • Cybenchの飽和（全問正解）、CyberGym（83.1%） • AISI（英国AI安全研究所）によるThe Last Ones（TLO）を初めて突破

[Anthropic 2026a][Anthropic 2026b][Anthropic 2026c][AISI 2026b]

Mythosだけではない

前頁で挙げたブレークスルーはMythosに限ったものではない。CyberGymのリザルトでも、Mythosだけがとびぬけているわけではない。
また、OpenAIのDaybreakとGPT-5.5-Cyberによる自律型ワークフローなどは、Mythosと同様に「自律型エージェント」に移行している。
つまり、AIを利用したセキュリティが、新しいフェーズに入ったと考える必要がある。

Leaderboard

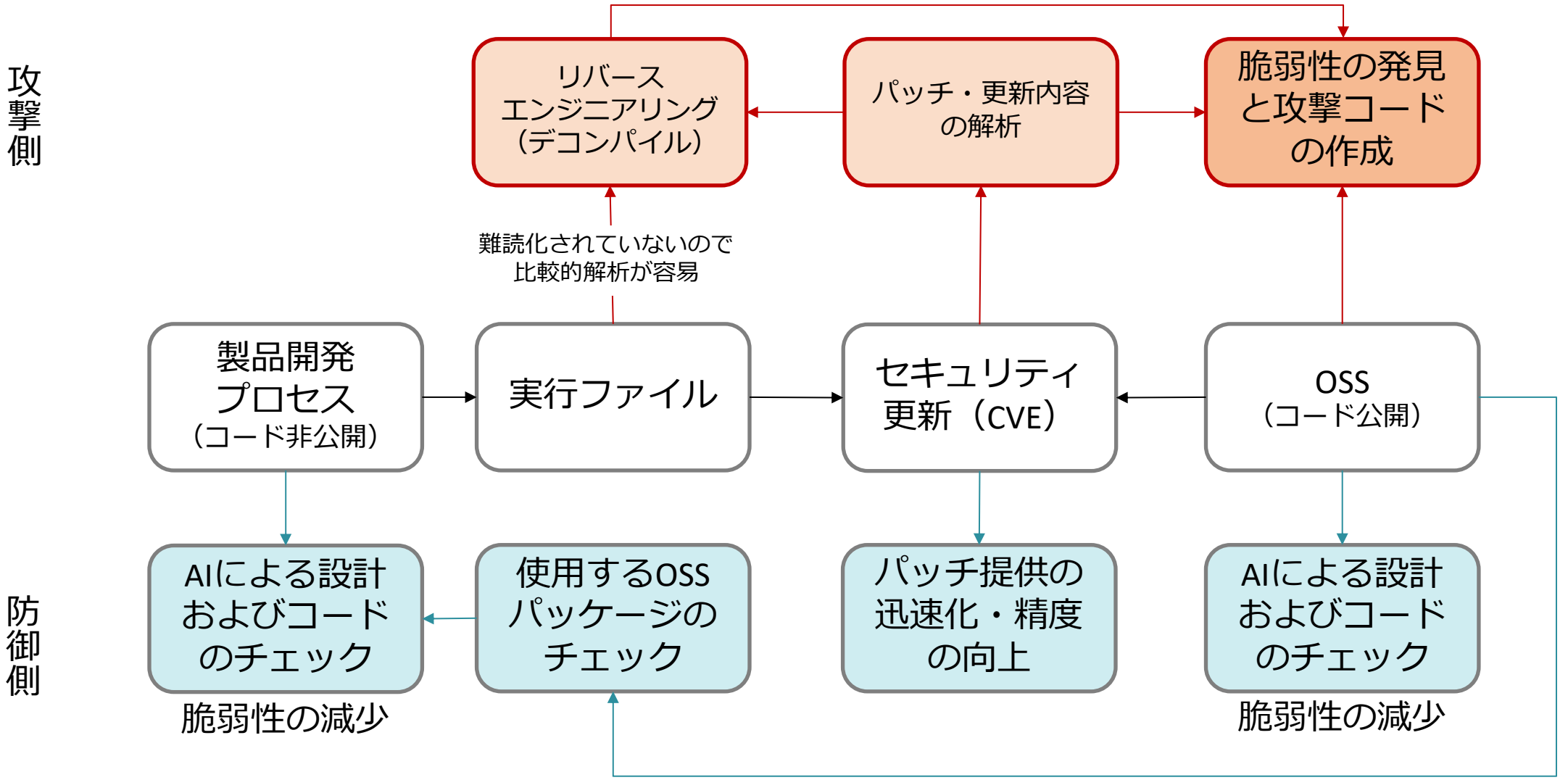
Model	Unguided % Solved	Subtask-Guided % Solved	Subtasks % Solved	Most Difficult Task Solved (First Solve Time by Humans)	
				Unguided	Subtask-Guided
Claude Mythos Preview [6]	100%	--	--	--	--
Claude Opus 4.7 [8]	96%	--	--	--	--
Claude Opus 4.6 [5]	93%	--	--	--	--
Claude Opus 4.5 [3]	82%	--	--	--	--
Muse Spark [7]	65.4%	--	--	--	--
Claude Sonnet 4.5 [3]	60%	--	--	--	--
Grok 4 [4]	43%	--	--	--	--
Claude Opus 4.1 [3]	42%	--	--	--	--
Grok 4.1 Thinking [4]	39%	--	--	--	--
Claude Opus 4 [2]	38%	--	--	--	--
Claude Sonnet 4 [2]	35%	--	--	--	--
Grok 4 Fast [4]	30%	--	--	--	--
OpenAI o3-mini [1†]	22.5%	--	--	42 min	--
Claude 3.7 Sonnet [1]	20%	--	--	11 min	--

Cybench

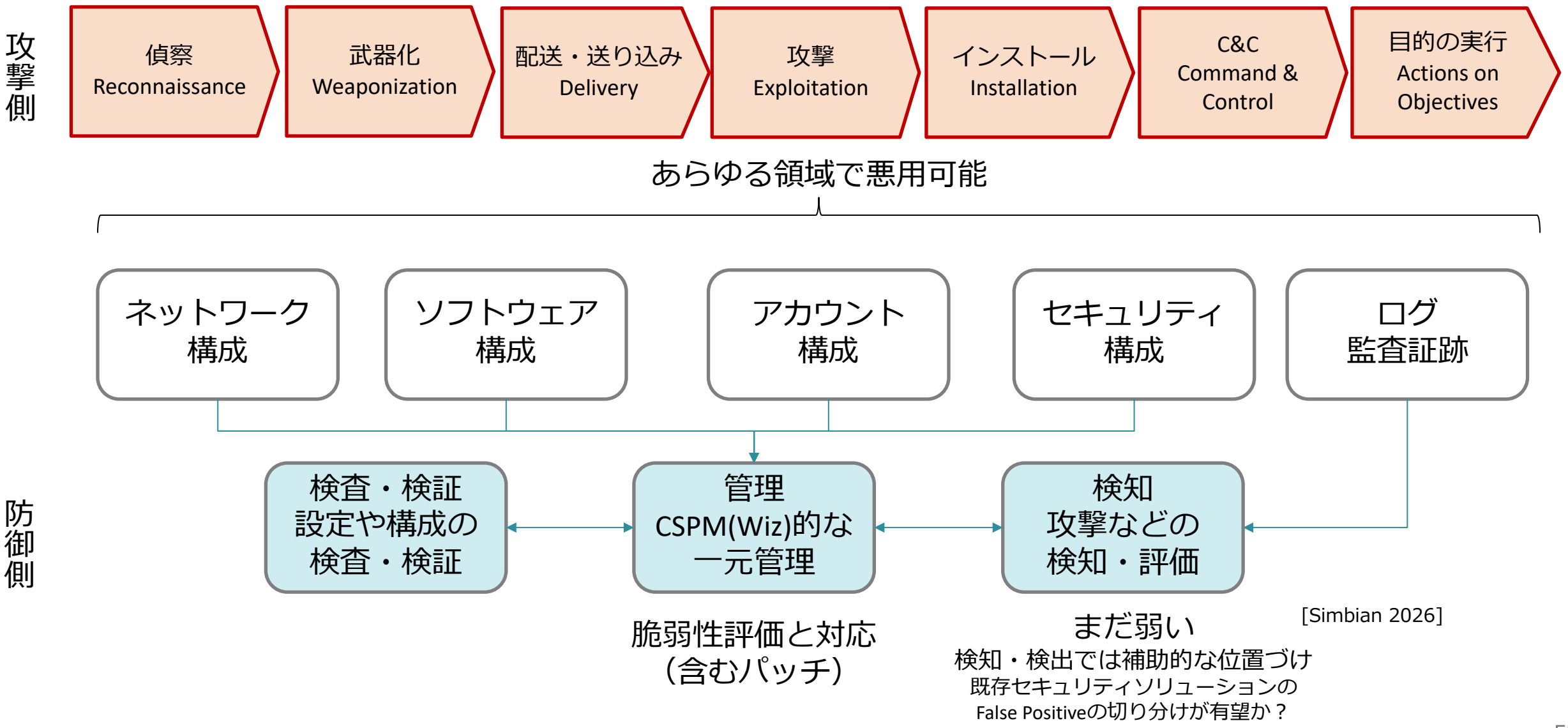
CyberGym

それぞれ（攻撃・防御）に想定されるAI利用のメリット

AIの位置づけ：脆弱性と攻撃コード



AIの位置づけ：運用環境



AIの業務利用について（AIリテラシー）

AIを活用しないことがリスクになる

AIを使うと「AIを使わないことが最大のリスク」と納得できる

- AIはよく働く
  - AIチャットボットでも、調査だけではなく、企画の策定や検証にとても有益
    - 画像イメージをテキスト化するだけでも結構便利
    - AIを利用することは、いつでも使えるアソシエートを雇ったような感じ
      - しかも、勝手に中断したり、要求を変更しても問題にならない
    - 思ったように動かないときは、自分の説明能力を反省することになる

- 「AIなんて結局、役に立たない」と思ったら…
  - 基本的なプロンプトのセオリーは一通り学んだ方が良い
  - AIチャットに特性があることも考慮する
    - 同じサービスでも、複数のモデルや、複数のサービスが選択できる。
    - モデルによって得手不得手があるので、複数のサービスを試せる環境が望ましい

責任と権限は利用者にある

- 機密情報の入力に気を遣うよりは安全なサービスを選ぶ
  - まずは、安全なサービスを選ぶこと
  - そのうえで、情報クラスに従った運用を行う

- AIは嘘をつくというか、「無理に答えようとする」
  - 新人に仕事を頼んだとか、アソシエートの限界はこんなもの、などと思えばよさそう
  - 責任は自分にあることを自覚し、重要な情報は必ず裏をとる
    - 事例:Deloitte AIチャットボット誤情報事例

- AIはバイアスを強める傾向がある
  - AIが質問に沿った応答するため、利用者の考えを補強し、偏った認識につながることがある
    - 品質はこうあるべきだとか、セキュリティ施策はこうあるべきだ、など自分で経験したので怖さがわかるけど、自覚するまで気づきにくい
      - 極端な例：「ChatGPTが自殺の原因に」　4人の遺族がオープンAIを提訴

[Ledge.ai 2024] [朝日新聞 2025]

AI Agent(コード生成) のパラダイムシフト

セキュリティのパラダイムシフトかも

CISOが本当に懸念すべきは…

- AI コード生成ツールが、パラダイムシフトになると思う
  - 環境を構築し、動くコードセット作り、バグを修正し、環境を変えることができる
  - AIがコードだけではなくシステム・構成を理解する
  - UNIXやPCが出た時の衝撃に近い、しかもインターネットも一緒に来た感じ

- 何が起きるか
  - 能力を拡張し（実現できなかったスキルの獲得）IT計画への自由度が一気に拡大する
    - エンジニアリングスキルはあった方が良いに決まっているが、
      スキルが無くても公開システムまで作れてしまう。
    - 既存のシステムへの機能追加や仕様変更が容易になる。
  - IT技術者に求められるスキルが変わる
    - 企画力、設計能力、プロジェクトマネジメント能力が重要になる

CISOが本当に懸念すべきは…

- そして、どうなる？
  - 基礎的な知識や、経験のない人が作ったシステムが増える
  - ノーコードというか、コードを見ないシステム開発が蔓延する
    - しかも存在を認識することが難しい
  - かつて統制なく広まった個人作成の業務ツール群（EUC*）の悪夢がよみがえる
    - 誰もメンテナンスできない神EXCEL, 担当者が退職したAccess DB、出自のわからないマクロ機能などをメンテナンスできなくなった問題と同じ構図が起こりうる

- 何が問題になるか
  - AIに作らせたシステムが、セキュリティ上の懸念点となる
    - 統制が取れない、インベントリ管理もできなくなる（もちろんSBOMを構築できない）
    - セキュリティアップデートなどのメンテナンスが行われない
    - 設計上の脆弱性が散在する
    - 責任者（オーナー）があやふやになる
    - Tokenなどの認証情報の漏洩が頻発
  - 未統制システムの増殖リスク

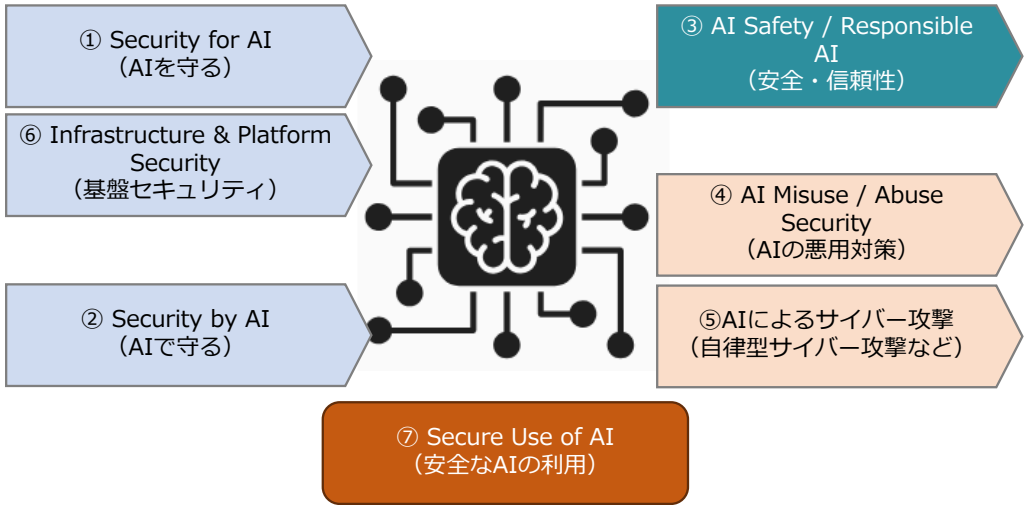
- なにが必要か
  - 「AIシステムへの懸念と対応」を参照ください

*EUC: End User Computing

むすび

CISOにとってのAIシステム

- AIのセキュリティは、まだまだ不明瞭
  - AI安全性
  - AI特有のセキュリティ
  - サイバーセキュリティ
  - AI品質



- AI安全性が中心にAIセキュリティが議論されているが…
  - AI安全性は、もちろん大事だが…
  - AIシステムやAIで作成したシステムは
    サイバーセキュリティ上の問題を数多く内包する

  - CISOは自覚的にこのリスクに対応する必要がある

AIがCISOの能力を拡張する

CISOがAIを使わない限り、本質的な理解を得ることはできない。
AIシステムの利便性と脅威をCISO自らが理解する必要がある。

AIがCISOの能力を拡張する

- 技術的な視点
  - 部下に頼むか外注が必要だったツールを自分で作れるようになる
    - 省力化、自動化、など
  - セキュリティの強化にも利用できる（脆弱性チェック、システム構成のチェックなど）
- 調査
  - 規格・ガイドライン、製品、手法などを調べるときに、基礎的な調査をまとめさせることができる（特に、Deep Research, Researchなど）
- 報告書
  - ファクトに基づいた報告書作成が容易になる
  - 文章の校正を行うことで、つまらないミスを避けることができる
- リスク評価
  - 専門外のリスク評価の端緒を得ることができる

参考文献

- [OpenAI 2018] Improving Language Understanding by Generative Pre-Training, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- [OpenAI 2022] ChatGPT が登場, https://openai.com/ja-JP/index/chatgpt/?utm_source=chatgpt.com
- [PPC 2023] 個人情報保護委員会生成 AI サービスの利用に関する注意喚起等について, https://www.ppc.go.jp/files/pdf/230602_kouhou_houdou.pdf
- [Mitchell+ 2019] Model Cards for Model Reporting, https://arxiv.org/abs/1810.03993
- [Google 2025] Gemini 2.5 Pro Model Card, https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Pro-Model-Card.pdf
- [OpenAI 2025] OpenAI GPT-5 System Card, https://arxiv.org/abs/2601.03267
- [Anthropic 2025] System Card: Claude Opus 4.5, https://www.anthropic.com/claude-opus-4-5-system-card
- [AISI 2025] AISI AIセーフティ評価のための評価ツール, https://aisi.go.jp/output/output_information/250912/
- [AC 2025] 国立情報学研究所　大規模言語モデル研究開発センター AnswerCarefully Dataset, https://llmc.nii.ac.jp/answercarefully-dataset/
- [OWASP 2024] OWASP Top 10 for LLMs and Gen AI Apps 2023-24, https://genai.owasp.org/llm-top-10-2023-24/
- [QA4AI 2024] AIプロダクト品質保証コンソーシアム AIプロダクト品質保証ガイドライン https://raw.githubusercontent.com/qa4ai/Guidelines/refs/heads/main/QA4AI.Guidelines.202504.pdf
- [OKDT 2026] claude-code-hardening-cheatsheet https://github.com/okdt/claude-code-hardening-cheatsheet

参考文献

- [METI/MIC 2025] 経済産業省 AI事業者ガイドライン,
  https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20240419_report.html
- [AISI 2026] AISI　AI-IRS AI インシデントレスポンス・アプローチブック
  https://aisi.go.jp/assets/pdf/ai-irs_v1.0_ja.pdf
- [TREND 2026a]　OpenClawの不正なスキルを通してmacOS型情報窃取ツール「AMOS」が拡散
  https://www.trendmicro.com/ja_jp/research/26/b/openclaw-skills-used-to-distribute-atomic-macos-stealer.html
- [TREND 2026b] Trend Micro社 AIゲートウェイがバックドアに：LiteLLMサプライチェーン侵害の内幕
  https://www.trendmicro.com/ja_jp/research/26/d/inside-litellm-supply-chain-compromise.html
- [時事 2026] 時事通信 政府、重要インフラの防御強化　１５分野、ミュトス対策を決定！
  https://www.jiji.com/jc/article?k=2026051800606&g=pol
- [日経 2026a] 日本経済新聞 3メガバンク、AIミュトスのアクセス権入手へ　サイバー防衛で日米連携
  https://www.nikkei.com/article/DGXZQOUB130SI0T10C26A5000000/
- [JBpress 2026] 小林啓倫 目前に迫る"バグマゲドン"、Appleが5年の歳月をかけて構築した　最強セキュリティOSを5日で破ったMythosの脅威 JBpress, 2026.5.18,
  https://jbpress.ismedia.jp/articles/-/94870
- [日経XTECH 2026] 日経クロステック Mythosが引き起こす「脆弱性の嵐」に備えよ、　専門家250人が対策を提言
  https://xtech.nikkei.com/atcl/nxt/column/18/00676/050700224/
- [Anthropic 2026a] Anthropic System Card: Claude Mythos Preview
  https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf
- [Anthropic 2026b] Carlini N., et al. Assessing Claude Mythos Preview's cybersecurity　capabilities, Anthropic Frontier Red Team
  https://red.anthropic.com/2026/mythos-preview/
- [AISI 2026b] AI Security Institute Our evaluation of Claude Mythos Preview's cyber　capabilities
  https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities
- [Anthropic 2026c] Anthropic Project Glasswing: Securing critical software for the AI era
  https://www.anthropic.com/glasswing
- [Ledge.ai 2024]デロイトの報告書に生成AIのハルシネーションで存在しない文献を引用・参照、豪政府に代金を一部返金――脚注誤りを訂正し再公開、コンサル業界に波紋
  https://ledge.ai/articles/deloitte_ai_refund_australia_report
- [朝日新聞 2025]「ChatGPTが自殺の原因に」　4人の遺族がオープンAIを提訴
  https://www.asahi.com/articles/ASTC836J6TC8UHBI012M.html
- [Simbian 2026] Cyber Defense Benchmark:Agentic Threat Hunting Evaluation for LLMs in SecOps
  https://arxiv.org/pdf/2604.19533


Thank you