CISOがおさえておきたい AIセキュリティの基本

Ver.02 2026-08

JNSA CISO支援ワーキンググループ
高橋　正和

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、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

Ⅰ -1 Security for AI
（AIを守る）

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

Ⅱ Security by AI
（AIで守る）

Ⅲ AI Safety / Responsible AI
（安全・信頼性）

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

Ⅳ-2 AIによるサイバー攻撃
（自律型サイバー攻撃など）

Ⅴ 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 Ⅴ. 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

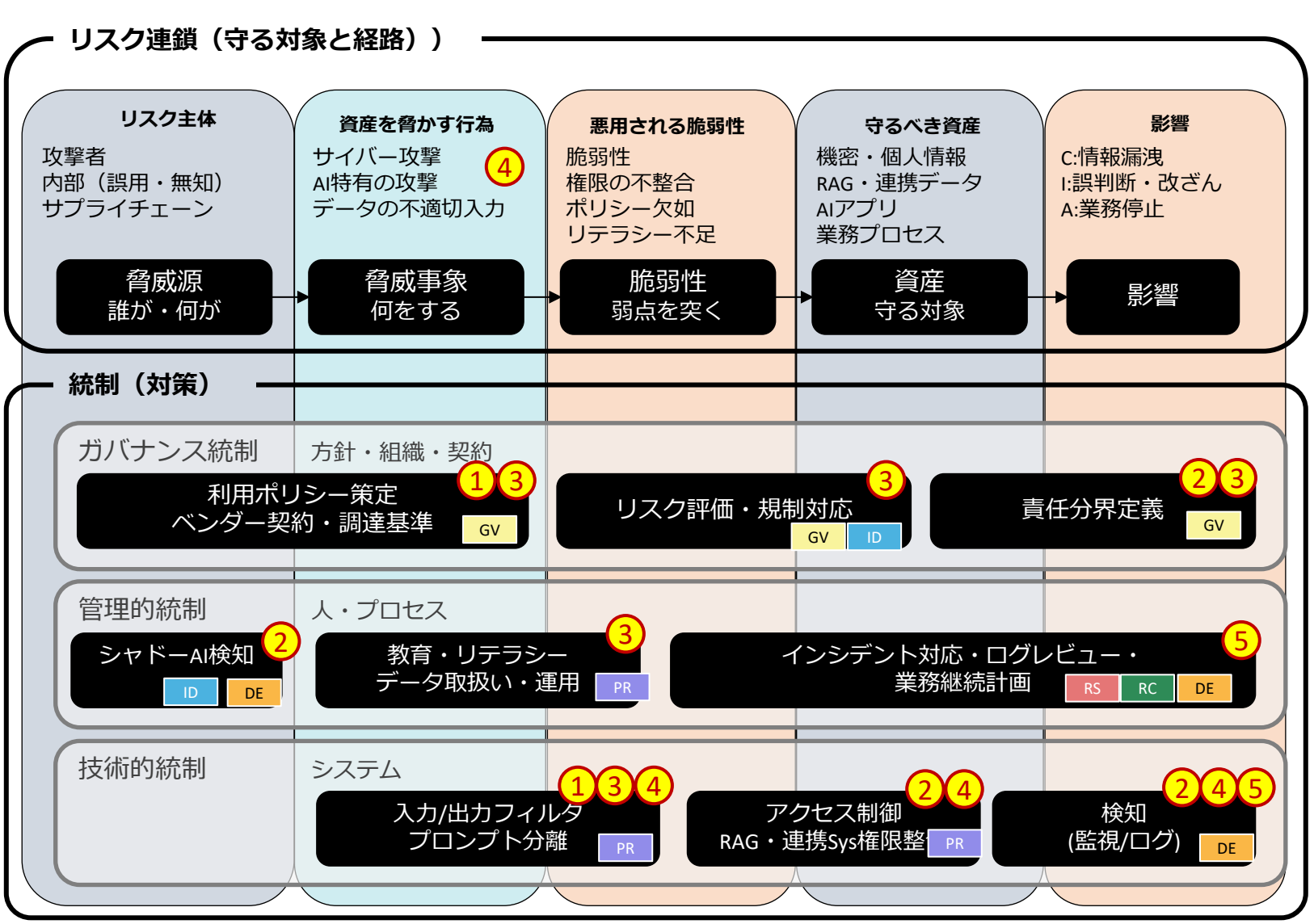
①の前ですが…まずは

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

… だけじゃない

AI利用の懸念点に対する施策の構成 +CSF



① データ・データ入力に係る懸念
② IT統制上の懸念
③ システムの応答に係る懸念
④ AIシステムへのサイバー攻撃への懸念

データクラスの実装 1 2

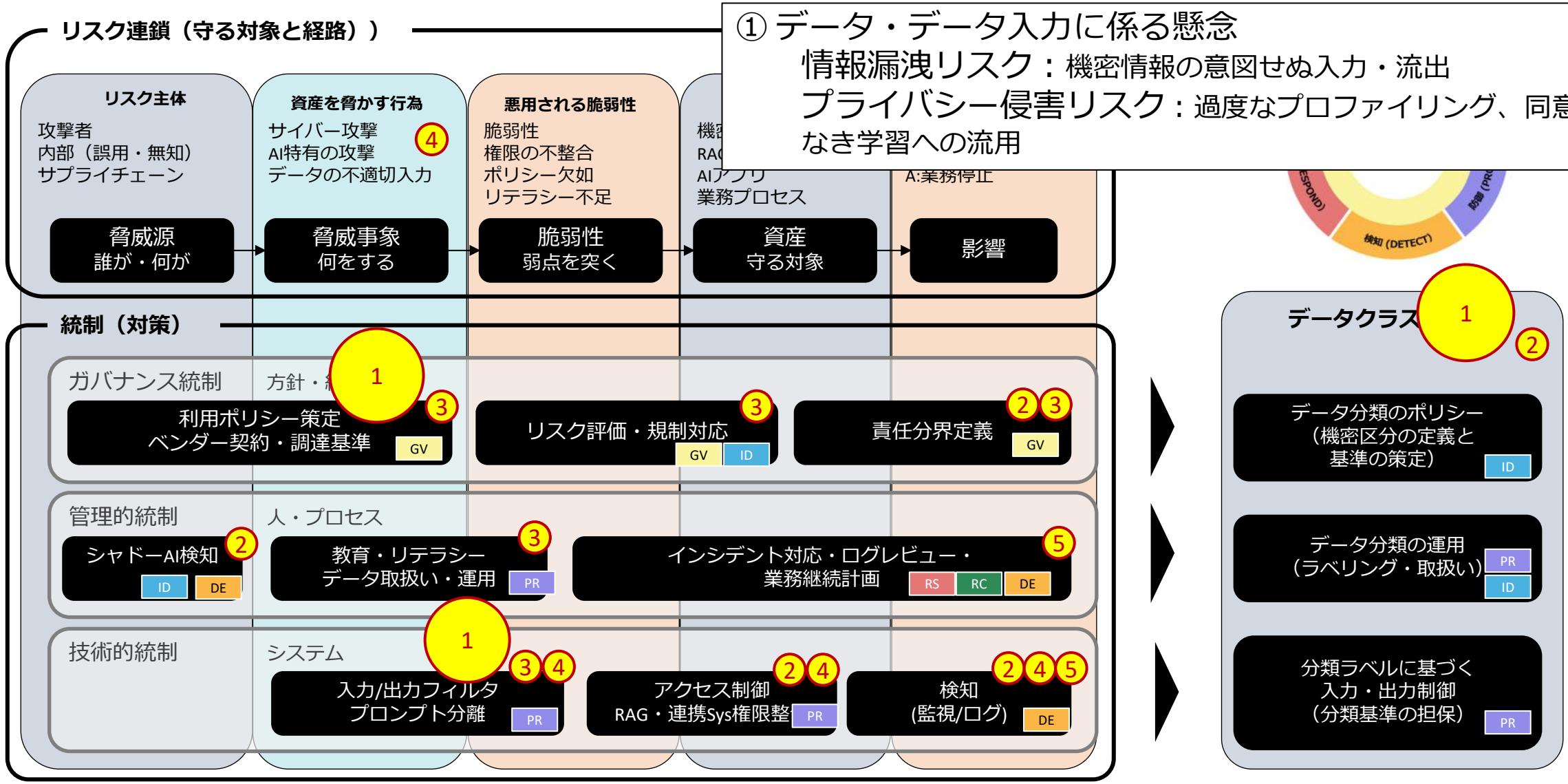
- データ分類のポリシー（機密区分の定義と基準の策定） ID
- データ分類の運用（ラベリング・取扱い） PR ID
- 分類ラベルに基づく入力・出力制御（分類基準の担保） PR

CSFタグ　GV 統治　ID 識別　PR 防御　DE 検知　RS 対応　RC 復旧

+CSFとして[NIST CSF]をリファレンス

AI利用の懸念点に対する施策の構成 +CSF

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

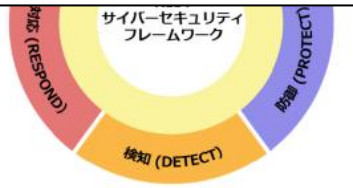


CSFタグ　GV 統治　ID 識別　PR 防御　DE 検知　RS 対応　RC 復旧

+CSFとして[NIST CSF]をリファレンス

AI利用の懸念点に対する施策の構成 +CSF

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



リスク連鎖（守る対象と経路））

リスク主体	資産を脅かす行為	悪用される脆弱性		
攻撃者 内部（誤用・無知） サプライチェーン	サイバー攻撃 ④ AI特有の攻撃 データの不適切入力	脆弱性 権限の不整合 ポリシー欠如 リテラシー不足	RAG・連携データ AIアプリ 業務プロセス	I:誤判断・改ざん A:業務停止
脅威源 誰が・何が	脅威事象 何をする	脆弱性 弱点を突く	資産 守る対象	影響

統制（対策）

	方針・組織・契約	人・プロセス	システム
ガバナンス統制	利用ポリシー策定 ベンダー契約・調達基準 ①③ GV リスク評価・規制対応 ③ GV ID 責任分界定義 ②③ GV		
管理的統制		シャドーAI検知 ② ID DE 教育・リテラシー データ取扱い・運用 ③ PR インシデント対応・ログレビュー・ 業務継続計画 ⑤ RS RC DE	
技術的統制			入力/出力フィルタ プロンプト分離 ①③④ PR アクセス制御 RAG・連携Sys権限整 ②④ PR 検知 (監視/ログ) ②④⑤ DE

データクラスの実装 ①②

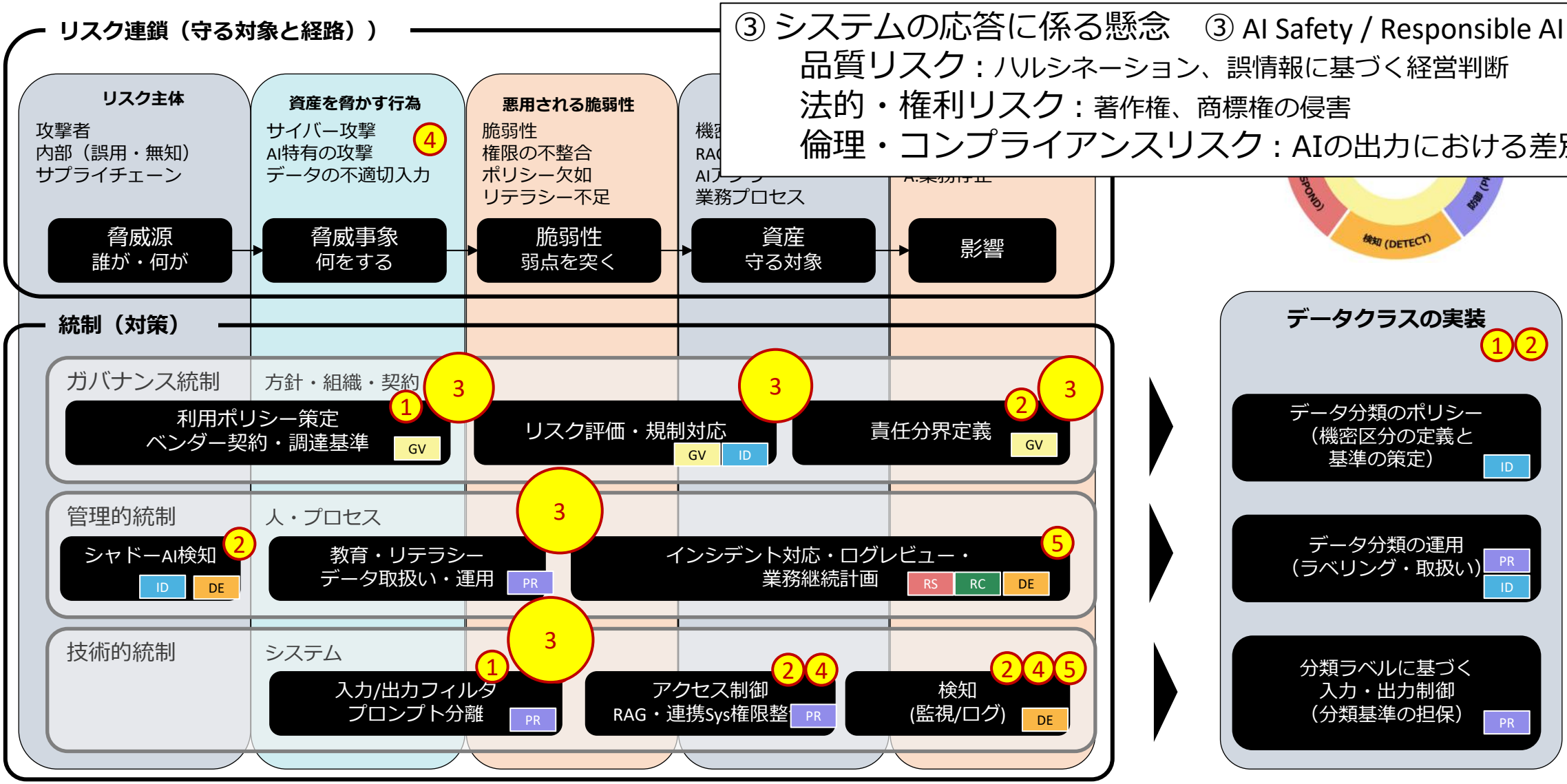
- データ分類のポリシー（機密区分の定義と基準の策定） ID
- データ分類の運用（ラベリング・取扱い） PR ID
- 分類ラベルに基づく入力・出力制御（分類基準の担保） PR

CSFタグ　GV 統治　ID 識別　PR 防御　DE 検知　RS 対応　RC 復旧

+CSFとして[NIST CSF]をリファレンス

AI利用の懸念点に対する施策の構成 +CSF

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

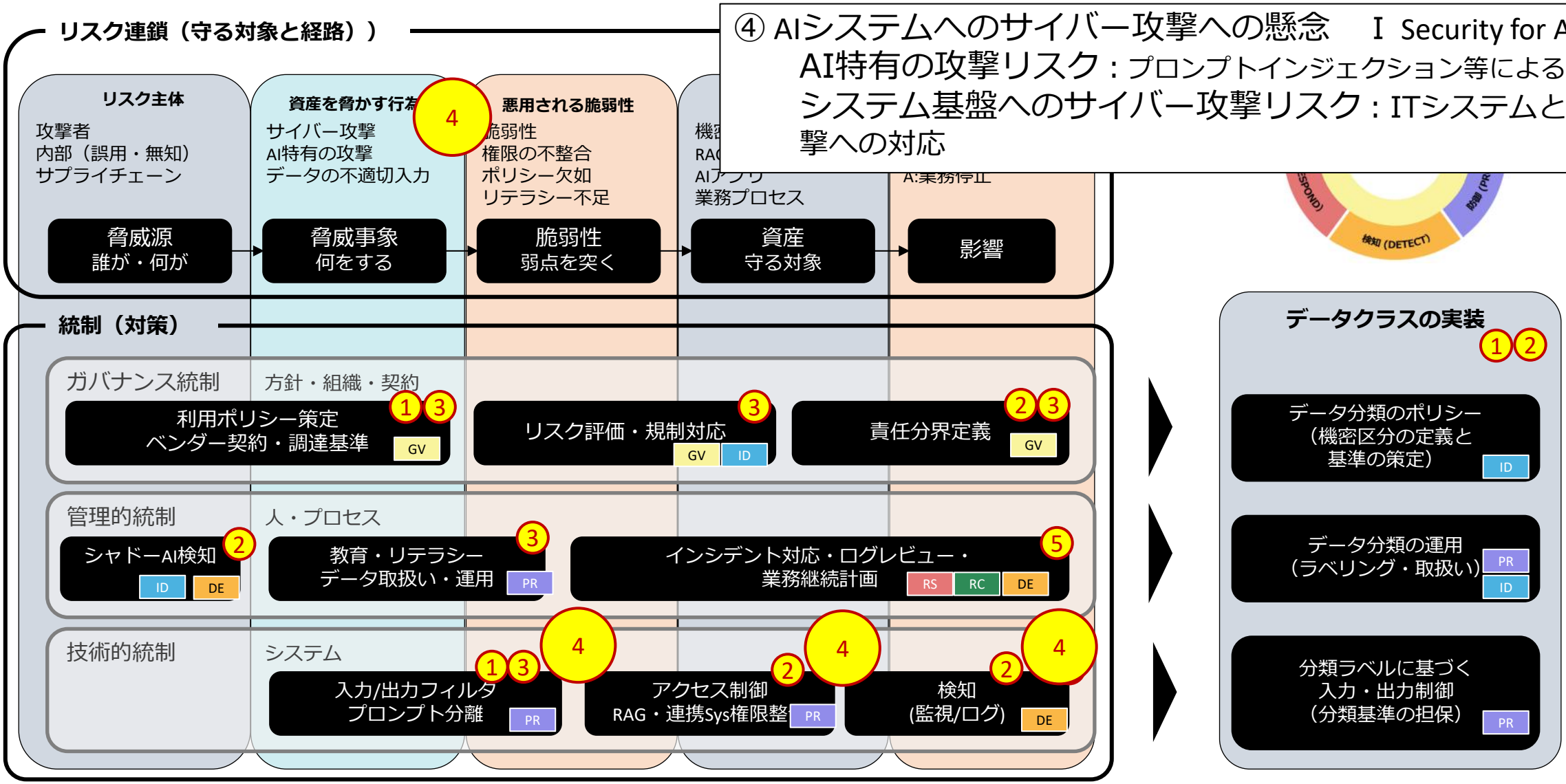


CSFタグ　GV 統治　ID 識別　PR 防御　DE 検知　RS 対応　RC 復旧

+CSFとして[NIST CSF]をリファレンス

AI利用の懸念点に対する施策の構成 +CSF

④ AIシステムへのサイバー攻撃への懸念　Ⅰ Security for AI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

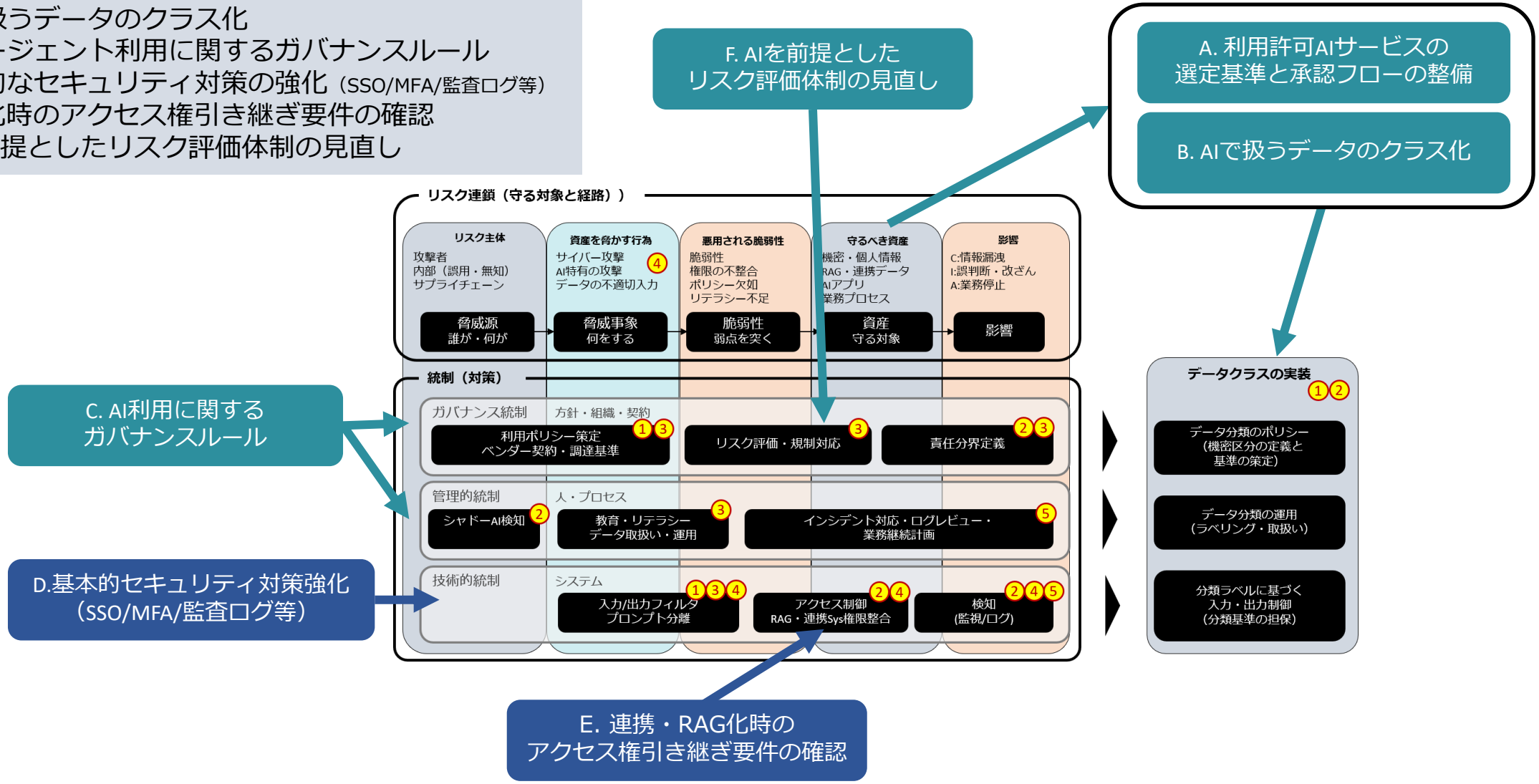


CSFタグ　GV 統治　ID 識別　PR 防御　DE 検知　RS 対応　RC 復旧

+CSFとして[NIST CSF]をリファレンス

C. AIに係る脅威と施策のマッピング

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. 基本的なセキュリティ対策の強化（SSO/MFA/監査ログ等）
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



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

情報漏洩とプライバシーの懸念を整理する

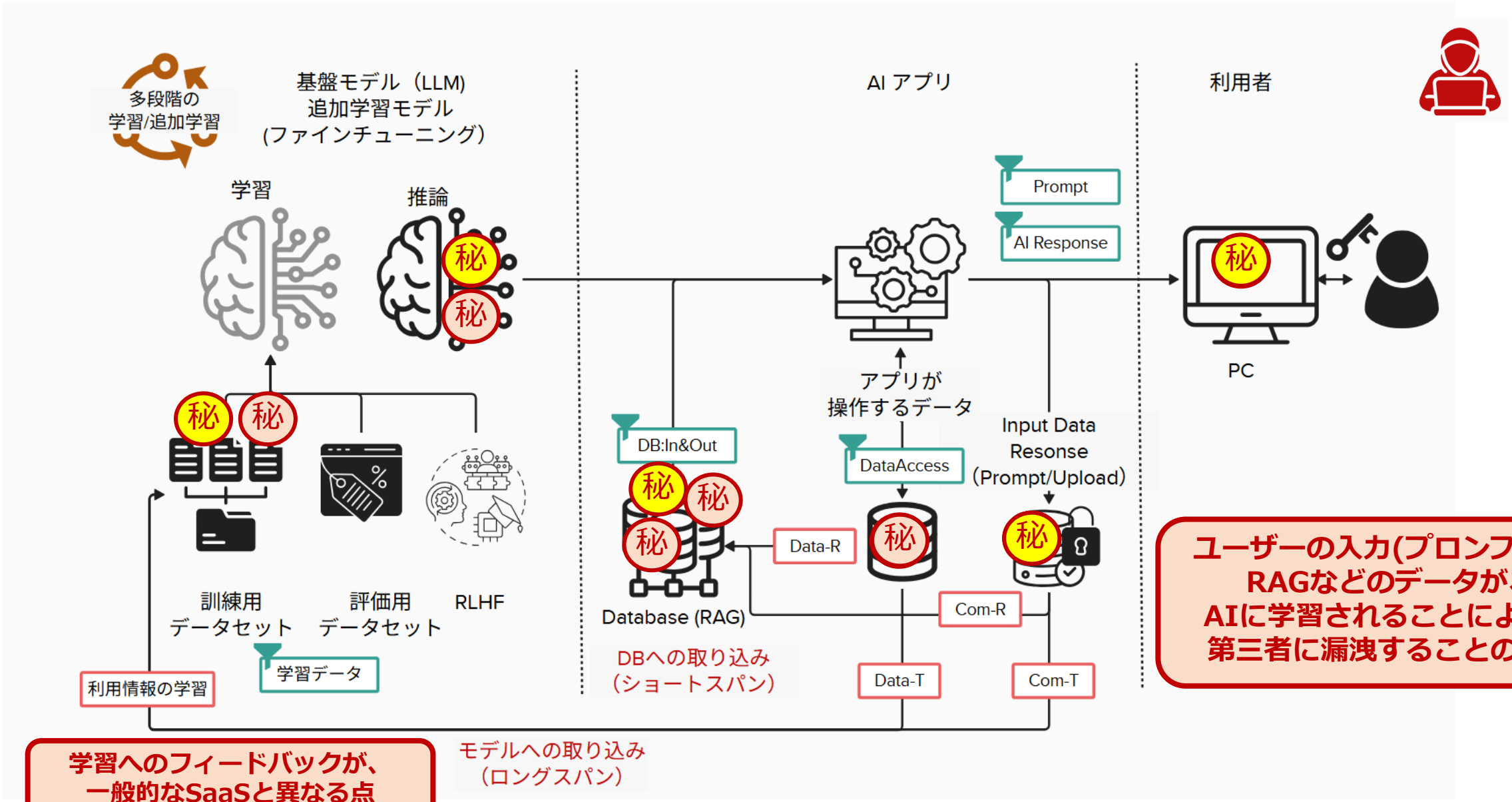
① データ・データ入力に係る懸念

- 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

AIアプリのユーザーデータ利用モデル



**ユーザーの入力(プロンプト)や
RAGなどのデータが、
AIに学習されることにより、
第三者に漏洩することの懸念**

**学習へのフィードバックが、
一般的なSaaSと異なる点**

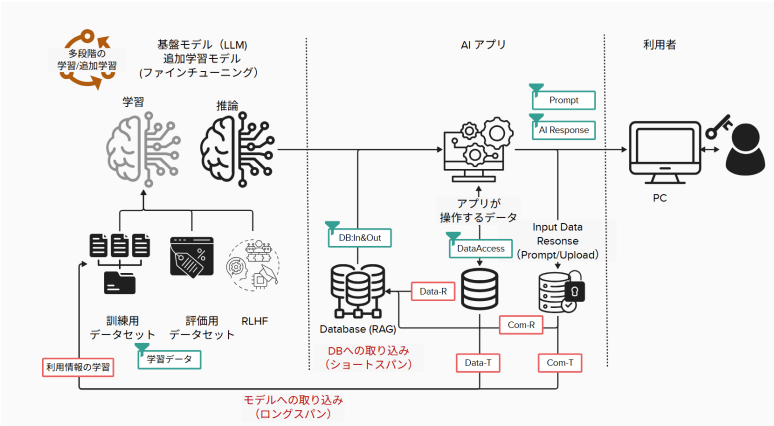
ユーザーデータの学習の有無

Free版の初期画面

Enterprise版のチャット画面

OpenAI　2026-08-22に取得

サービス毎のユーザーデータの学習の有無



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。

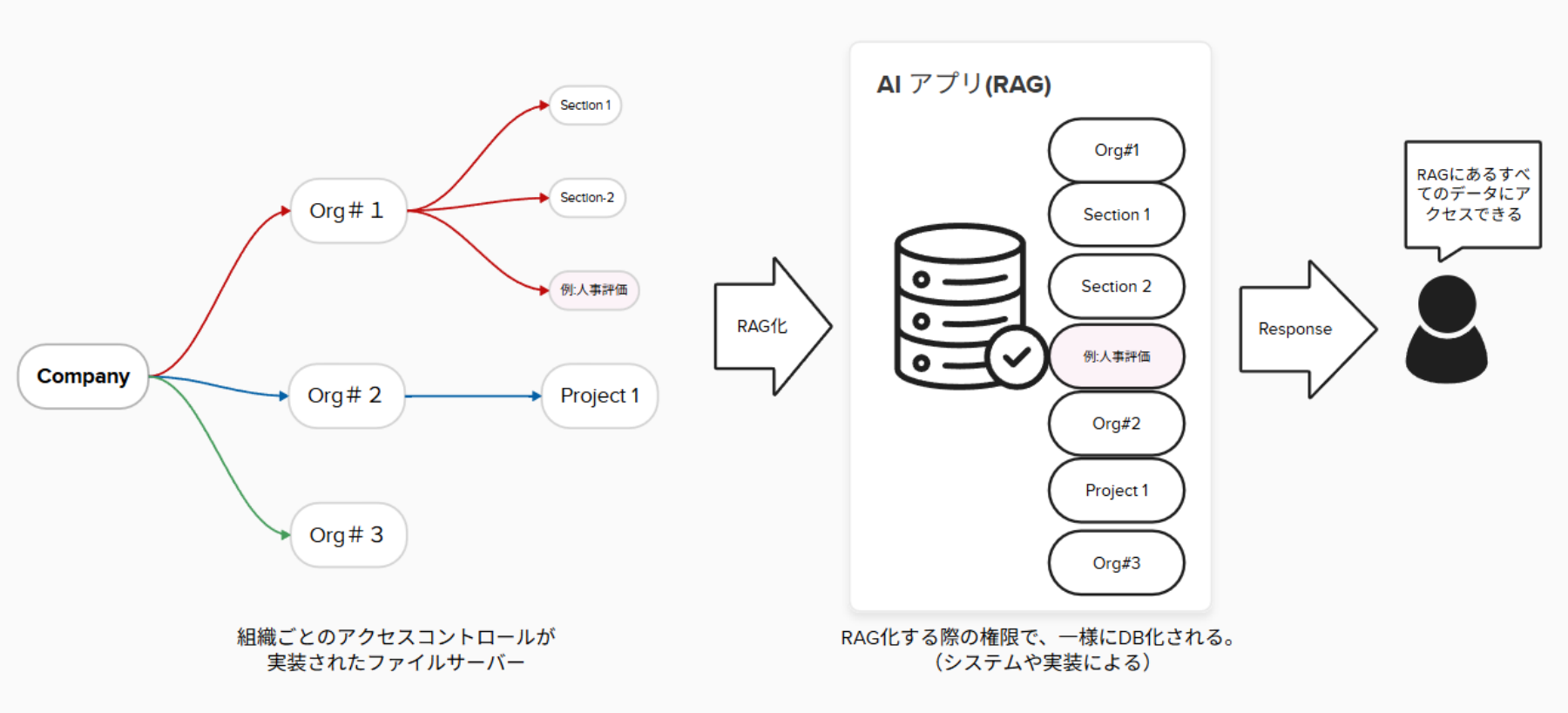
シャドーAIの利用は、入力情報が学習データに利用されることに伴って、当該AIの他の利用者に漏洩する可能性があり、広範囲な情報漏洩につながることが懸念される。

	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	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	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	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	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	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	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

E. RAG化に伴う権限情報の剥落リスク

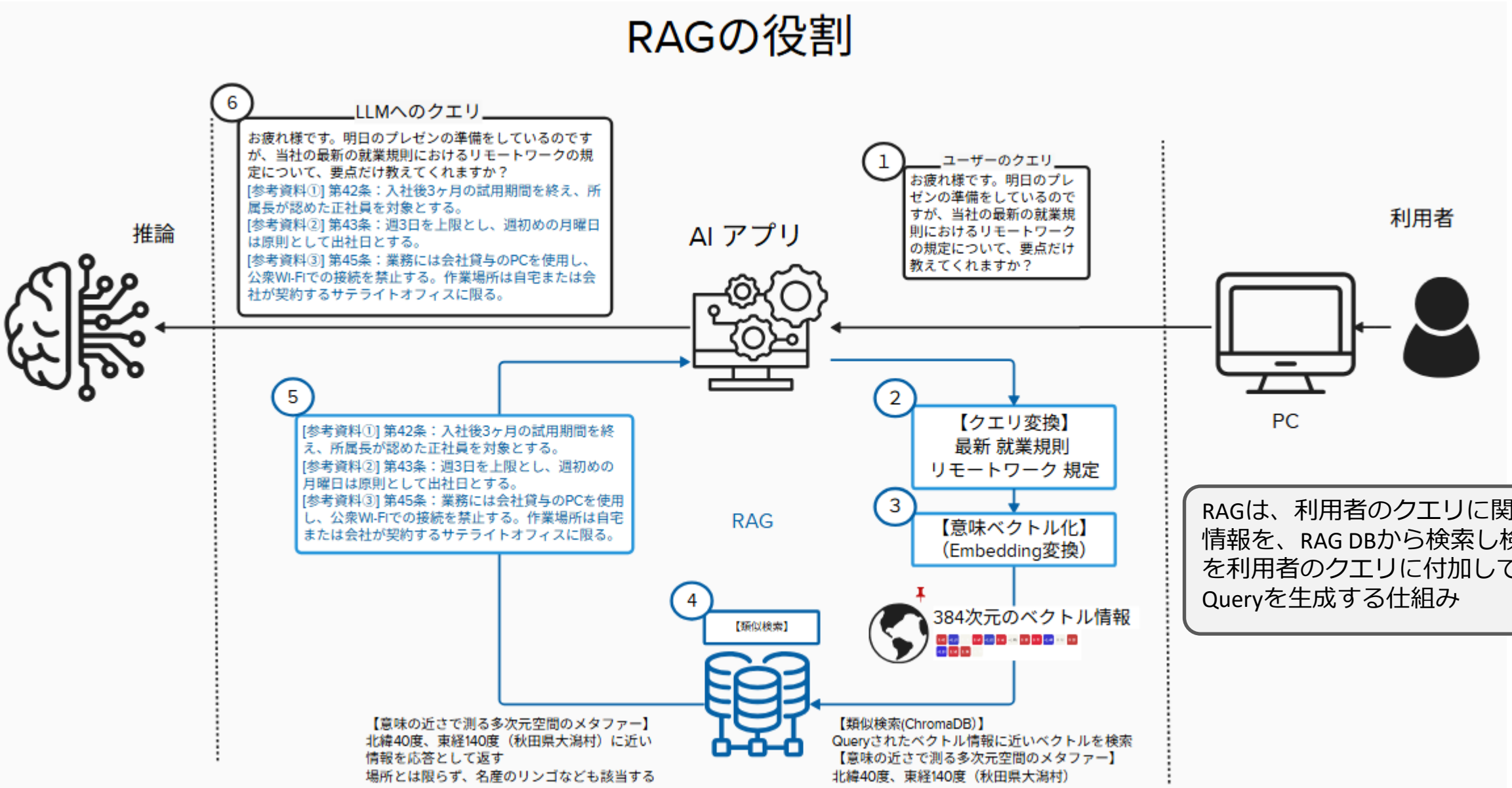
ファイルサーバーのように、アクセスコントロールが行われているデータを
RAG化すると、アクセスコントロールが失われる可能性がある。
例えば、社内ドキュメントのAI利用には、アクセス権剥落への注意が必要。
（データソースの権限管理を、RAGでも実装できるか確認が必要）



組織ごとのアクセスコントロールが
実装されたファイルサーバー

RAG化する際の権限で、一様にDB化される。
（システムや実装による）

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

【対策】
学習による
情報漏洩・プライバシー侵害

A.B. データクラスと手続きの構成例

データクラスの実装

データ分類のポリシー
（機密区分の定義と基準の策定）
CIS 3.7データ分類スキーム

データ分類の運用
（ラベリング・取扱い）
CIS 3.1 データ管理プロセス
3.4 データの保持
14.4 データ取扱いの従業員教育

ラベルに基づく入力・出力制御
（分類基準の担保）
CIS 3.3 データアクセス制御リスト
3.12 機密度に応じた処理・保管の分離
3.13 データ盗難防止（DLP）
3.14 機密データへのアクセス記録

A. 利用許可AIサービスの選定基準と承認フローの整備

- 自社の調達チェックリストを整備する。第三者認証の有無、再委託、越境の有無、データの権利（削除・返却条件）、適用法令（開示請求リスク等）などの要件を明確にする
- セキュリティ要件を整備する。特に入力データを学習に使わせない（オプトアウト）、出力の取り扱い（権利帰属・二次利用の可否・第三者権利侵害時の補償）を盛り込む
- 確認の実務：法務・情報システム・データ保護責任者を必ず承認フローに巻き込む

B. AIで扱うデータのクラス化

区分	クラウド／AI利用の可否（要点）
極秘 (Restricted)	営業秘密・未公開情報・特定個人情報など。明示的に許可されたものを除き外部AI利用は不可、閉域・専用テナント＋契約での保証が必須
秘 (Confidential)	第三者認証（ISO 27001/27017、SOC2 Type II、ISMAP等）取得が前提 契約での学習除外・データ削除権、技術的な保護の明記を必須とする AI固有の管理についてはISO/IEC 42001の取得状況の確認が望ましい
社内限 (Internal)	利用可(アクセス制御・暗号化)。ただし画一的な約款同意のみのサービスでは秘以上の情報は扱わない
公開 (Public)	利用可

C. AI利用に関するガバナンスルール

1. **オプトアウトの徹底** —— 入力した要機密情報を学習させない仕様か
2. **データ所在国の確認** ——海外サーバは現地法令・開示請求のリスク
3. **提供元基準** —— 単独では個人情報でなくても、照合で個人特定できれば個人情報

③ システムの応答に係る懸念

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

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

ハルシネーションが問題となった例

- チャットボットの誤った回答が問題となったケース
  - チャットボットが航空券の払い戻しについて誤った回答をした事案
    ［American Bar Association, 2024］
- 納品物にAIのハルシネーションが含まれていたケース
  - 政府への納品物にAIが生成したハルシネーションが含まれていた
    ［Business Standard, 2025］
- 弁護士がChatGPTに生成させた訴状に、実在しない判例6件を引用
  - [Mata v. Avianca, 2023]
- Google検索「AI Overviews」の風刺サイトに基づいた誤回答が拡散
  - 「石を食べると健康に良い」「ピザのチーズが剥がれないよう接着剤を混ぜるとよい」
    [Forbes, 2024]

法的・権利リスクの例 1/2

- 類型1：学習データの無断利用（著作権　開発・学習段階）
  - Andrea Bartzらが、約48万点の作品を対象とし、Anthropicが書籍をClaudeの学習に使ったことを争った。[Bartz et al. v. Anthropic, 2025]

- 類型2：出力物による権利侵害（著作権・商標　生成・出力段階）
  - Getty Imagesは、Stability AIが1,200万点超の写真・キャプション・メタデータを無断利用してStable Diffusionを構築したとして、2023年に提訴し、英国では2025年11月に一部商標権侵害のみ認容する判決が出た。[Getty Images v. Stability AI, 2025]

- 類型3：報道コンテンツの利用と要約サービス（著作権＋不正競争）
  - 2023年、NYTはMicrosoftとOpenAIに対し、記事を無断で学習に用い、自社報道の「市場代替」を構築しているとして、著作権侵害に加え、不正競争および商標希釈化を主張し、数十億ドルの損害賠償を求めた。（不正競争の主張は2025年3月に、商標希釈化の主張は2026年6月に取り下げ・却下されている）。 [The New York Times v. Microsoft & OpenAI, 2023]

法的・権利リスクの例 2/e

- 類型4：AI生成物の著作権帰属（著作物性・人間の創作性）
  - AIシステムが自律生成した美術作品の登録を拒否した米国著作権局に対し、司法審査を求めた事案。2026年3月2日、連邦最高裁は上告を受理せず（cert. denied）、著作物には人間による著作者性が必要とするD.C.巡回区控訴裁判所の判断が確定した。[Thaler v. Perlmutter, 2026]

- 類型5：声・肖像などパブリシティ／隣接権
  - 2024年5月、声優2名が、「研究目的のみ」等して提供したサンプルが、AI音声クローンとして無断で複製・販売されたと主張。著作権法・商標法に基づく主張の大部分は認めなかったが、声の不正流用については、州民事権利法（パブリシティ権）に基づく主張が成立しうるとされ、州消費者保護法違反および契約違反の主張も認められた。 [Lehrman & Sage v. Lovo, 2025]

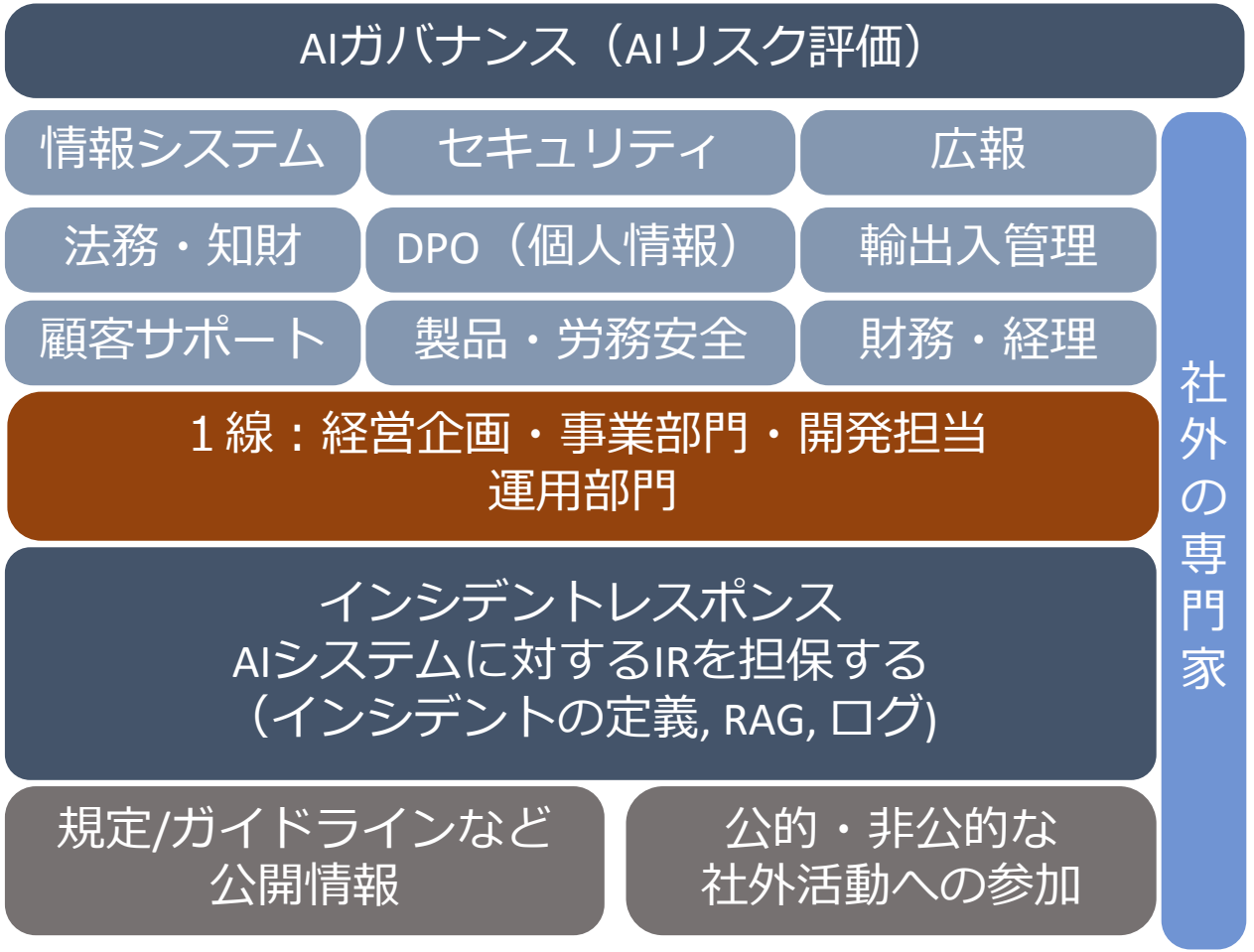
【対策】
システムの応答に係る懸念

F. AIを前提としたリスク評価体制

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？

社内外のステークフォルダーと連携して対応する



OECDのインシデント定義

潜在的な被害	実際に発生した被害
重大なAIハザード (a) 個人の死亡、または個人もしくは集団の健康に対する重大な危害 (b) 重要インフラの管理・運用に対する深刻かつ不可逆的な障害 (c) 基本的人権の重大な侵害、あるいは基本的権利・労働権・知的財産権を保護するために適用される法律に基づく義務の重大な違反 (d) 財産・地域社会・環境に対する深刻な損害 (e) 地域社会や社会の機能不全を引き起こし、当該地域が自らの資源を用いて対処する能力を試したり超えたりする可能性のある事態	**AIディザスター** AIインシデントの中でも特に深刻な者であり、コミュニティや社会の機能を阻害し、その独自の資源を用いた処理能力を試したり飛び越えたりする可能性のある事象を指す。AIディザスターの影響は即時的かつ局所的なものもあれば、広範囲に及び長時間持続するものもある。 **重大なAIインシデント** (a) 人の死亡、または人もしくは集団の健康に対する重大な被害 (b) 重要インフラの管理・運用に対する深刻かつ不可逆的な障害 (c) 基本的人権の重大な侵害、または基本的人権・労働権・知的財産権を保護するための 適用法令上の義務の重大な違反 (d) 財産・地域社会・環境に対する深刻な被害
AIハザード (a) 個人または集団の身体的損傷または健康被害 (b) 重要インフラの管理・運用の混乱 (c) 基本的人権の侵害、または基本的人権・労働権・知的財産権を保護するための適用法令上の義務違反 (d) 財産・地域社会・環境への損害	**AIインシデント** (a) 個人または集団の健康に対する傷害または被害 (b) 重要インフラの管理・運用の混乱 (c) 基本的人権の侵害、あるいは基本的人権・労働権・知的財産権を保護するための適用 法令に基づく義務違反 (d) 財産・地域社会・環境に対する被害

[OECD 2026]

④ AIシステムへの サイバー攻撃の懸念

- 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の特有のセキュリティとサイバーセキュリティ

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

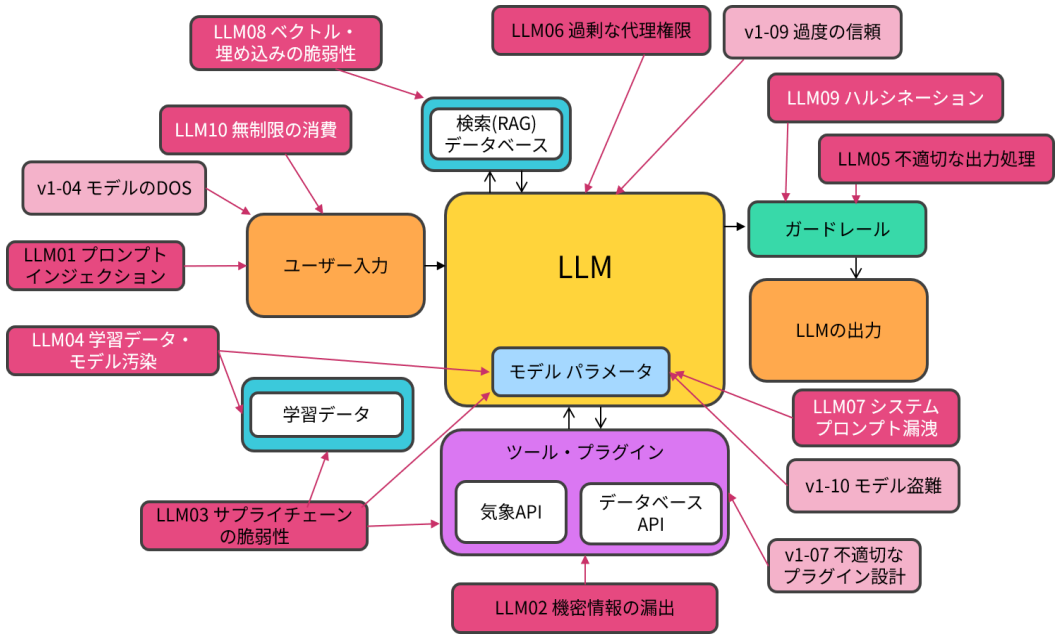
- 必要な対策
  - A. 利用許可AIサービスの選定基準と承認フローの整備　①で記載
  - B. AIで扱うデータのクラス化　①で記載
  - C. AIエージェント利用に関するガバナンスルール　①②で記載
  - D. SSO/MFA/監査ログ等のセキュリティの基本的な強化
    - AIサービスに対するSSO/MFAの実装
    - AI関連のインシデント対応体制の構築
    - AIエージェントで作成するアプリ・ツールのインベントリ
  - E. RAG化時のアクセス権引き継ぎ要件の確認　①で記載
  - F. システム開発のリスク評価体制の構築　③で記載

OWASP TOP10 for LLM Applications

- OWASPが公表する、LLMを組み込んだアプリケーション固有のセキュリティリスクを重要度順に10項目へ整理したガイドライン

#	脅威名	主な影響
LLM01:2025	プロンプト・インジェクション	不正操作・情報漏洩・システム侵害
LLM02:2025	機密情報の漏出	個人情報漏洩・知的財産侵害
LLM03:2025	サプライチェーン	バックドア・完全性の毀損
LLM04:2025	データ・モデル汚染	誤動作・バイアス・バックドア
LLM05:2025	不適切な出力処理	XSS・SQLインジェクション・RCE
LLM06:2025	過剰な代理行為	意図しない自律的アクション
LLM07:2025	システムプロンプト漏洩	攻撃の足がかり・機密漏洩
LLM08:2025	ベクトル・埋め込みの脆弱性	RAG汚染・情報操作
LLM09:2025	誤情報（ハルシネーション）	意思決定エラー・法的リスク
LLM10:2025	使用制限の不備（無制限の消費）	DoS・経済的損害・モデル窃取

[OWASP LLM , 2025]

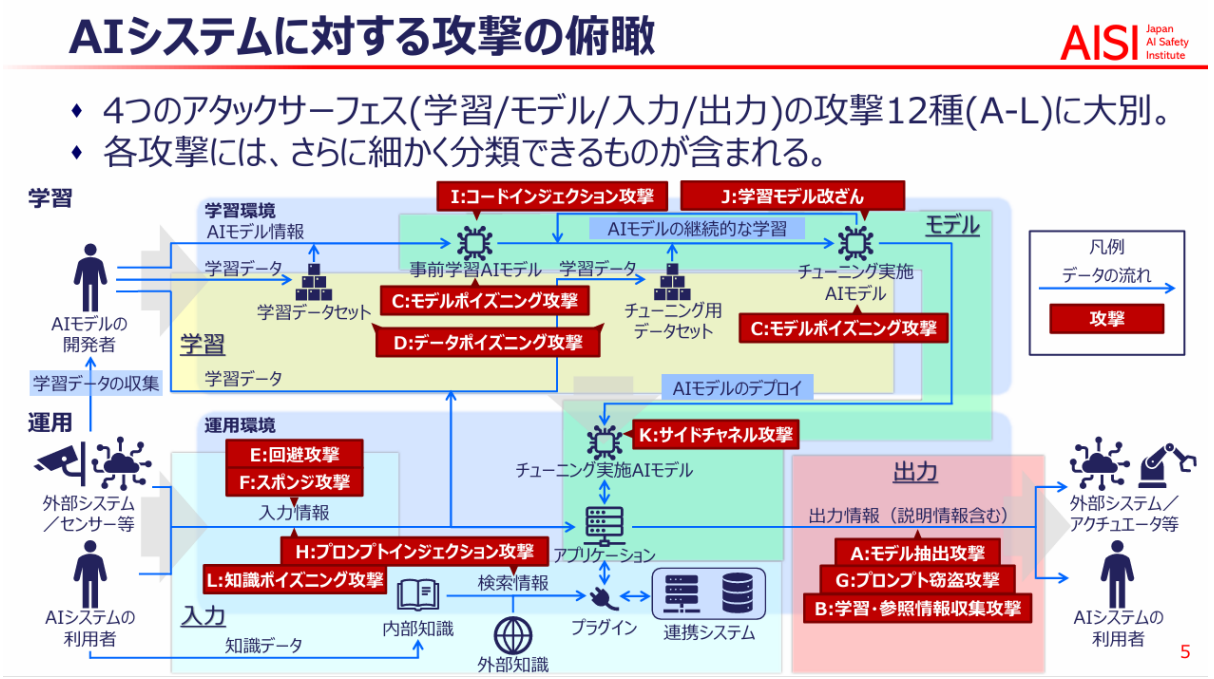


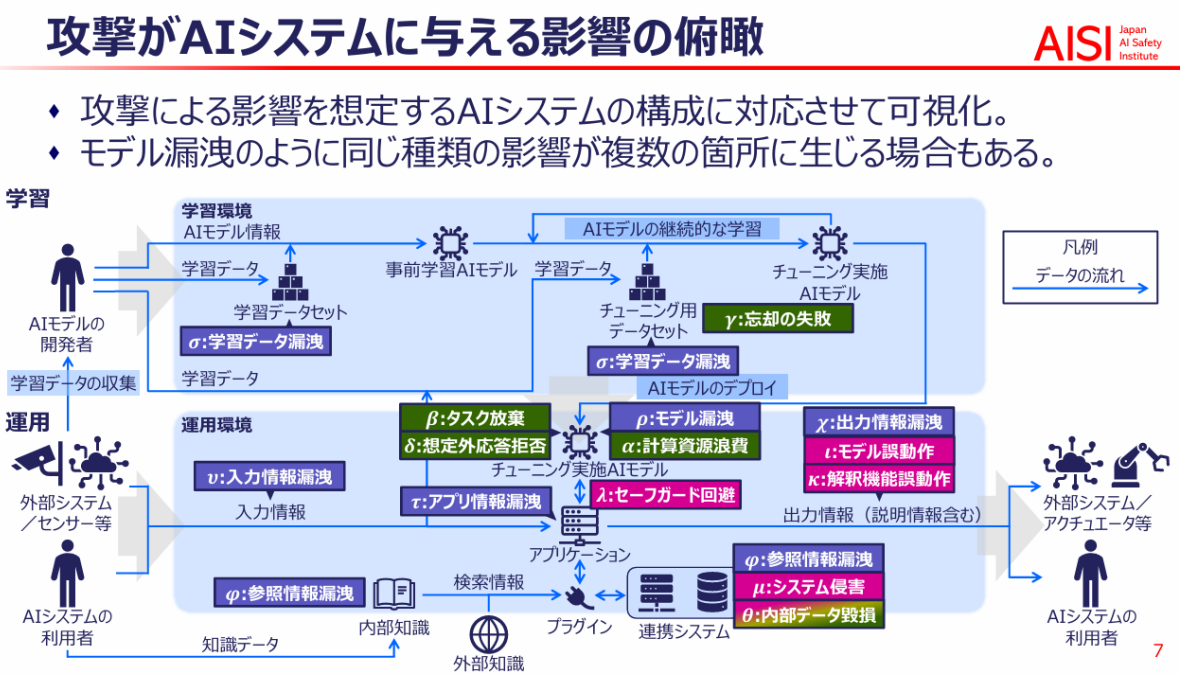
[Confident AI 2025]を筆者が修正・翻訳

気象API:読み込み専用のAPIの例、データベースAPI:書き込みを伴うAPIの例

日本AISI AIシステムに対する既知の攻撃と影響

- 日本AISI「AIシステムに対する既知の攻撃と影響」は、「AIシステムに対する特有のセキュリティ攻撃を俯瞰すべく、学術論文等で発表されたAIやAIシステムに対する攻撃とその影響をまとめた資料 」
- AIシステムに対する攻撃を俯瞰し、攻撃がAIシステムに与える影響を分類・整理した資料となっており、リファレンスも丁寧に記載されている。





[JP AISI KAI, 2026]

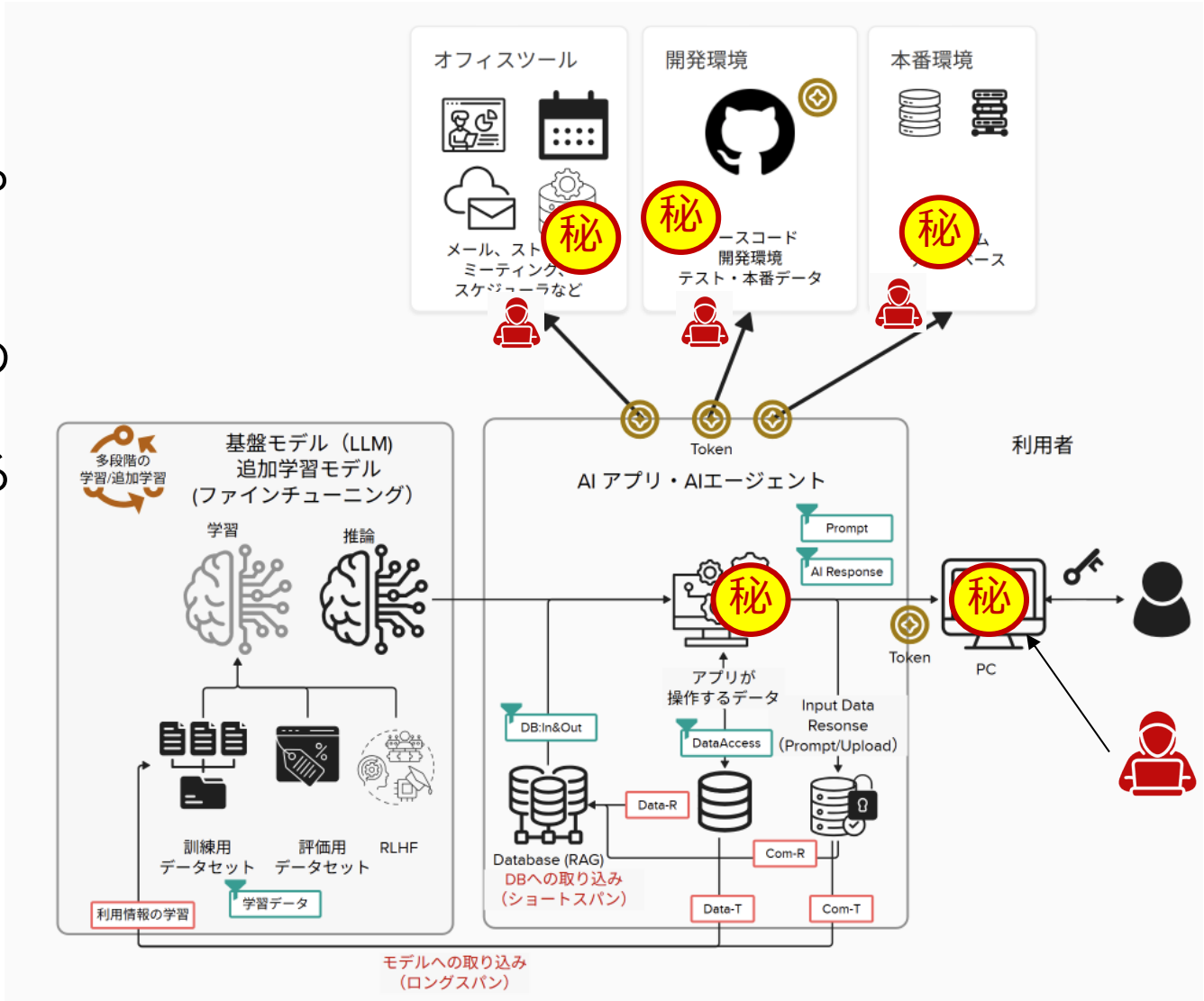
【対策】
AIシステムへの
サイバー攻撃への懸念

D. セキュリティの基本的な強化（ SSO/MFA ）

IAAA（認証・認可）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

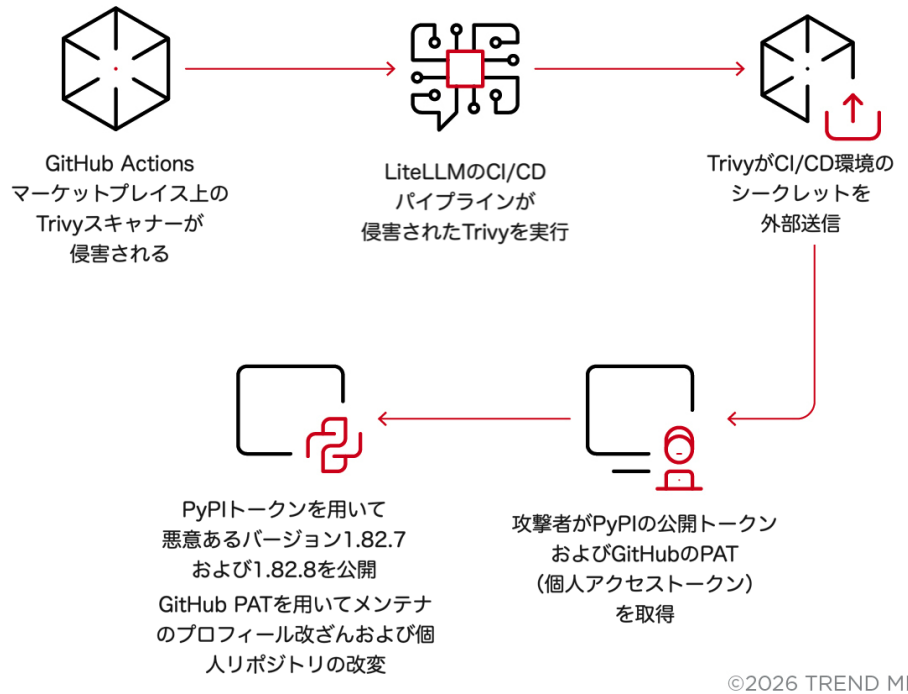
- 識別（Identification）
- 認証（Authentication）
- 認可（Authorization）
- 監査（Accountability/Audit）

認証情報の重要性：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 2026]



Mythos Previewの衝撃

2026年4月に公表されたMythos Previewは、
　　脆弱性検出、PoCの作成、自律的な攻撃ができるとされた。
セキュリティ施策を根底から変えてしまうという衝撃が走った

Mythosのブレークスルー
Ⅳ-2 AIによるサイバー攻撃（自律型サイバー攻撃など）の懸念

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。

Discovery
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）

Exploitation
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）

Intrusion
自律的な攻撃
（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][UK AISI 2026b]

CyberGym @2026-08-23

2026-06-17

2026-08-23

CyberGym @2026-08-23

#	Agent / Model	Success Rate	Date	Source
1	Sangfor AI AGENT dynamic uses DeepSeek-V4-Flash	93.2%	2026-08-08	Sangfor AI ↗
2	Whitzard (白泽) AGENT dynamic uses DeepSeek-V4-Flash	91.2%	2026-08-07	Fudan Whitzard ↗
3	MDASH AGENT uses GPT-5.4, Claude Opus 4.6, Claude Sonnet 4.6	91.0%	2026-06-17	Microsoft ↗
4	Wiz Atlas AGENT uses GPT-5.5, Claude Opus 4.6	90.9%	2026-07-27	Wiz ↗
5	DoGNAVY AGENT dynamic uses GLM-5.2	90.8%	2026-08-03	deepsec@DARKNAVY ↗
6	Crystalline AGENT test-time mem. uses Claude Opus 4.6	89.6%	2026-06-08	Independent researcher ↗
7	RedbudAI AGENT dynamic test-time mem. uses GLM-5.2	86.3%	2026-08-10	Redbud ↗
8	GPT-5.5-Cyber MODEL OpenAI Agent	85.6%	2026-06-22	OpenAI ↗
9	Velldepth Agent AGENT uses XekRung	85.3%	2026-08-03	Alibaba Security ↗
10	Xuanwu Atuin AI AGENT dynamic uses GLM-5.2	84.8%	2026-07-22	Tencent Xuanwu Lab ↗
11	Claude Mythos Preview MODEL Anthropic Agent	83.1%	2026-04-07	Anthropic ↗
12	GPT-5.5 MODEL OpenAI Agent	81.8%	2026-04-23	OpenAI ↗
13	GPT-5.4 MODEL OpenAI Agent	79.0%	2026-04-23	OpenAI ↗

2026-08-23 現在では、
Claude Mythosは11位になっている。

特定のAIモデルだけが脅威なわけではなく
Mythos Previewが示した変化は、
AI全般の懸念事項になっている。

今回は取り上げないが、焦点がハーネスに
移行しているといわれている。

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

まずは、ここから

データクラスの実装

データ分類のポリシー
（機密区分の定義と基準の策定）
CIS 3.7データ分類スキーム

データ分類の運用
（ラベリング・取扱い）
CIS 3.1 データ管理プロセス
3.4 データの保持
14.4 データ取扱いの従業員教育

ラベルに基づく入力・出力制御
（分類基準の担保）
CIS 3.3 データアクセス制御リスト
3.12 機密度に応じた処理・保管の分離
3.13 データ盗難防止（DLP）
3.14 機密データへのアクセス記録

A. 利用許可AIサービスの選定基準と承認フローの整備

- 自社の調達チェックリストを整備する。第三者認証の有無、再委託、越境の有無、データの権利（削除・返却条件）、適用法令（開示請求リスク等）などの要件を明確にする
- セキュリティ要件を整備する。特に入力データを学習に使わせない（オプトアウト）、出力の取り扱い（権利帰属・二次利用の可否・第三者権利侵害時の補償）を盛り込む
- 確認の実務：法務・情報システム・データ保護責任者を必ず承認フローに巻き込む

B. AIで扱うデータのクラス化

区分	クラウド／AI利用の可否（要点）
極秘 (Restricted)	営業秘密・未公開情報・特定個人情報など。明示的に許可されたものを除き外部AI利用は不可、閉域・専用テナント＋契約での保証が必須
秘 (Confidential)	第三者認証（ISO 27001/27017、SOC2 Type II、ISMAP等）取得が前提 契約での学習除外・データ削除権、技術的な保護の明記を必須とする AI固有の管理についてはISO/IEC 42001の取得状況の確認が望ましい
社内限 (Internal)	利用可(アクセス制御・暗号化)。ただし画一的な約款同意のみのサービスでは秘以上の情報は扱わない
公開 (Public)	利用可

C. AI利用に関するガバナンスルール

1. **オプトアウトの徹底** —— 入力した要機密情報を学習させない仕様か
2. **データ所在国の確認** ——海外サーバは現地法令・開示請求のリスク
3. **提供元基準** —— 単独では個人情報でなくても、照合で個人特定できれば個人情報

むすび

OWASP State of Agentic AI Security and Governance

ほとんどの組織は、自らのガバナンス能力を上回る速度でエージェントを導入している。
既存のプ ログラムに追加予算を割り当てるだけでは、このギャップを埋めることはできない。

３つの主要な知見　2026年のエージェント型技術環境全体を通じて得られたv2の発見事項

脅威はいまや現実のものである

プロンプトインジェクションは、エージェント型AIへの攻撃における主要な侵入経路である。エージェントのサプライチェーンは、理論上の懸念から実際の悪用へと、わずか数ヶ月で移行した。Top 10 Agenticのほぼすべての分類項目に、少なくとも1件の文書化されたインシデントが存在するに至っている。

→ 脅威分析（Threat Analysis）・実インシデント

セーフティとセキュリティは収斂する

セキュリティ上の被害は「エージェントに何をさせてしまうか」から生じ、セーフティ上の被害は「エージェント自身が何をできてしまうか」から生じる。高い自律性を持つエージェントにおいては、同じ設計判断が両方のリスクを生み、同じテレメトリ（遠隔計測）が両方を検知し、同じ封じ込め対応が両方を解消する。

→ AIセーフティ対AIセキュリティ

ガバナンスは実装の速度に追いつかなければならない

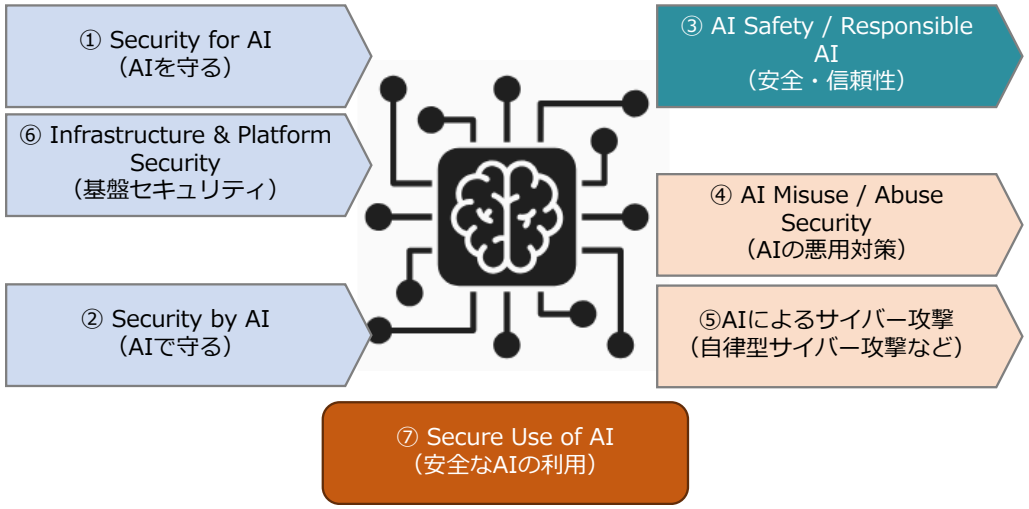
コーディングエージェントは今や日次でリリースを重ねており、新しいプロジェクトが数週間で企業利用の段階に達している。これは、従来のセキュリティレビューの体制が想定していた速度をはるかに超えている。四半期ごとのレビューや手作業でのトリアージでは、もはやこの速度に対応できず、異なる種類のツールが必要とされている。

→成熟度モデル・導入段階

[OWASP AASG 2026]

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


Thank you

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
- [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
- [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 2025] 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

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/pdf/20250328_1.pdf?utm_source=chatgpt.com
- [JP 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 2026] 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/
- [UK 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 2025]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

参考文献

- [American Bar Association, 2024] The American Bar Association (2024), BC Tribunal Confirms Companies Remain Liable for Information Provided by AI Chatbot , https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-february/bc-tribunal-confirms-companies-remain-liable-information-provided-ai-chatbot/
- [Business Standard, 2025] Deloitte's AI fiasco: Why chatbots hallucinate and who else got caught, https://www.business-standard.com/technology/tech-news/deloitte-ai-hallucination-report-australia-gpt4o-fabricated-references-125100800915_1.html
- [Mata v. Avianca, 2023] Mata v. Avianca, Inc. 判決文（Justia）
- [Forbes, 2024] Forbes: Google's AI Recommends Glue on Pizza, https://www.forbes.com/sites/jackkelly/2024/05/31/google-ai-glue-to-pizza-viral-blunders/
- [Bartz et al. v. Anthropic, 2025] Bartz et al. v. Anthropic PBC, Order on Fair Use, No. C 24-05417 WHA, Doc. 231, https://www.govinfo.gov/content/pkg/USCOURTS-cand-3_24-cv-05417/pdf/USCOURTS-cand-3_24-cv-05417-0.pdf
- [Getty Images v. Stability AI, 2025] Getty Images (US) Inc and others v Stability AI Ltd EWHC 2863 (Ch), 英国高等法院判決, https://www.judiciary.uk/wp-content/uploads/2025/11/Getty-Images-v-Stability-AI.pdf
- [The New York Times v. Microsoft & OpenAI, 2023] The New York Times Company v. Microsoft Corporation et al., No. 1:23-cv-11195, 訴状 https://www.courtlistener.com/docket/68117049/the-new-york-times-company-v-microsoft-corporation/
- [Thaler v. Perlmutter, 2026] Thaler v. Perlmutter, No. 25-449, cert. denied (U.S. Mar. 2, 2026) ; Thaler v. Perlmutter, 130 F.4th 1039 (D.C. Cir. 2025) ; The U.S. Copyright Office (2025), Copyright and Artificial Intelligence, Part 2: Copyrightability
- [Lehrman & Sage v. Lovo, 2025] Lehrman et al. v. Lovo, Inc., No. 1:24-cv-03770, Opinion, S.D.N.Y., July 10, 2025, https://copyrightalliance.org/wp-content/uploads/2025/07/Lehrman-v-Lovo-Opinion-July-10-2025.pdf

参考文献

- [OWASP LLM 2025] OWASP (2025), OWASP Top 10 for LLM Applications 2025 , https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/
- [Confident AI 2025] OWASP Top 10 2025 for LLM Applications: What's new? Risks, and Mitigation Techniques, https://www.confident-ai.com/blog/owasp-top-10-2025-for-llm-applications-risks-and-mitigation-techniques
- [OWASP ASI, 2026] Top 10 For Agentic Applications 2026, https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
- [OWASP AASG] OWASP GenAI Security Project（2026）、State of Agentic AI Security and Governance 2.01 https://genai.owasp.org/resource/state-of-agentic-ai-security-and-governance/
- [JP AISI KAI, 2026] AIシステムに対する既知の攻撃と影響, https://aisi.go.jp/output/output_security/known_attacks_and_impacts/
- [OECD 2026], OECD AI Incidents and Hazards Monitor (AIM), https://oecd.ai/en/incidents
- [NIST CSF] The NIST Cybersecurity Framework (CSF) 2.0, https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf
- [Conscia 2026] The OpenClaw security crisis, https://conscia.com/blog/the-openclaw-security-crisis/
- [ADVERSA 2026] OpenClaw security 101: Vulnerabilities & hardening , https://adversa.ai/blog/openclaw-security-101-vulnerabilities-hardening-2026/
- [Cybench] Cybench Leaderboard, Stanford University, https://cybench.github.io/#leaderboard_title
- [CyberGym 2026] CyberGym Leaderboard, UC Berkeley, https://www.cybergym.io/cybergym/