生成 AI および AI エージェントを
安全に活用するための手引書

2026 年 7 月

独立行政法人情報処理推進機構 産業サイバーセキュリティセンター

中核人材育成プログラム 9 期生

セキュリティ担当者のための生成 AI セキュリティ プロジェクト

改訂履歴

版数	改訂年月日	改訂箇所	改訂内容
第 1 版	2026 年 7 月 31 日		初版発行

目次

第 1 章　はじめに

1.1　背景

生成 AI および AI エージェントの急速な普及により、企業における生成 AI 導入・活用の動きは加速している。チャットボット、文書生成、コード補助、業務自動化など、生成 AI は既に多くの企業の日常業務に浸透しつつある。

こうした流れの中、政府や業界団体による生成 AI の安全な利用に向けたガイドラインも整備が進んでいる。例えば、総務省・経済産業省「AI 事業者ガイドライン[1]」は、AI 開発者・提供者・利用者といった各主体が取り組むべき指針や事項を整理した、AI ガバナンスの統一的な枠組みである。また、AI セーフティ・インスティテュート（AISI）「AI セーフティに関する評価観点ガイド[2]」は、主に AI システムの開発者・提供者を対象に、安全性評価を実施する際に参照すべき評価観点を体系的に示している。これらは、AI の開発から利用に至る各段階で考慮すべき事項や、安全性評価の観点を幅広く整理したものといえる。

一方で、これらの文書は、AI ガバナンスの基本的な考え方や評価の枠組みを示すことに主眼が置かれている。そのため、AI を自社業務に取り入れる利用者の立場に立ち、生成 AI を安全に導入・運用するための具体的な判断軸や、想定されるリスクの大きさに応じて実践的な対策を選び取るための指針を体系的に提供する文書は、いまだ十分に整備されているとはいえない。

また、企業内では「生成 AI は便利だが何となく怖い」という漠然とした懸念から導入が停滞する場合や[3]、組織が利用を許可していない AI システムを従業員が利用する（シャドーAI）問題についても、企業が AI 活用を進める上で看過できない課題であると考えられる。予算・リソース・ビジネススピードの制約が大きい企業にとって、大規模なガイドラインをそのまま適用することは現実的ではない。

本書はこうした背景のもと、企業が生成 AI を安全に導入・運用するために必要な知識と判断軸を、ライフサイクル全体を通じて体系的に提供する実践的な手引書である。

1.2　本書の目的

本書の目的は、AI システムのライフサイクルや、生成 AI のリスクをわかりやすく提示し、対策を提案することで、企業の安全な生成 AI の利活用を加速させることである。また、執筆にあたり、以下の問題意識を出発点としている。

---

[1] 総務省・経済産業省, AI 事業者ガイドライン（第 1.2 版）
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20260331_1.pdf

[2] AISI,AI セーフティに関する評価観点ガイド
https://aisi.go.jp/output/output_framework/guide_to_evaluation_perspective_on_ai_safety/

[3] 総務省,「令和 7 年版 情報通信白書」第 1 部第 1 章第 2 節「AI の爆発的な進展の動向」, 同白書では、生成 AI 導入に際しての懸念事項として、セキュリティリスク等が挙げられている。
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r07/html/nd112220.html

- 生成 AI の導入メリットを理解しつつも、不明確な脅威に対する漠然とした懸念から導入に慎重になる場合や、生成 AI のリスクを明確に理解しないまま活用が進む場合がある。
- 生成 AI のリスクを提示するのみでは、生成 AI 利活用の加速にはつながらない。さらに、理想的な対策を提示するのみでは相応のコスト・リソースを消費し、利活用へのハードルを過剰に引き上げてしまう。
- 産業界で安全な生成 AI 利活用を促進するためには、各企業が自らの生成 AI 利活用において明確な根拠を持ってビジネス影響を評価し、リスク管理方針（受容・低減など）およびリスク低減施策（技術、運用、ガバナンス）を判断することが必要である。

そのため本書は、企業が実業務や提供サービスの中で生成 AI を安全・安心して導入・活用できるよう、生成 AI の導入から運用・廃棄に至るライフサイクル全体を見据え、リスクと対策の判断に必要な情報を体系的に提供する。

1.3 本書における AI に関する用語の定義

本書が対象とする AI とは、「生成 AI」および「AI エージェント」であり、ロボット・自動運転等の「フィジカル AI」は本書の対象外とする。また、本書では、以下のように用語を定義する。

表 1-1 本書における AI に関する用語の定義

用語	定義
AI	生成 AI と AI エージェントを指す。
AI システム	活用の過程を通じて様々なレベルの自律性をもって動作し学習する機能を有するソフトウェアを要素として含むシステムを指す。（AI 事業者ガイドライン[4]の定義を引用） 上記定義する AI を構成要素に含む。
AI モデル	AI システムに含まれ、学習データを用いた機械学習によって得られるモデルで、入力データに応じた予測結果を生成する。（AI 事業者ガイドラインの定義を引用） LLM（大規模言語モデル）が代表例である。
生成 AI	文章、画像、プログラム等を生成できる AI モデルにもとづく AI の総称を指す。（AI 事業者ガイドラインの定義を引用）
AI エージェント	特定の目標を達成するために、環境を感知し自律的に行動する AI システムを指す。（AI 事業者ガイドラインの定義を引用）

■生成 AI と AI エージェントの境界

近年生成 AI のエージェント化が加速しており、生成 AI と AI エージェントの境界は曖昧になりつつある。例えば Gemini は生成 AI として提供されているが、Google の各種サービスと連携し、メールの自動

[4] 総務省・経済産業省, AI 事業者ガイドライン（第 1.2 版）
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20260331_1.pdf

作成やカレンダー操作といったエージェント機能を備えている。同様に ChatGPT もプラグインや GPTs 機能[5]（カスタム GPT）を通じて外部ツールと連携し、単なるテキスト生成を超えた自律的なタスク実行が可能となっている。

このように現在の主要な AI システムは、生成 AI を基盤としながらエージェント機能を取り込む方向へ進化しており、両者を厳密に区別することは実務上困難になりつつある。実務上重要なのは両者の区別よりも、AI システムが持つ自律性の程度である。自律性が高まるほど、AI は人間の介入なしに外部ツールの操作やデータへのアクセスを行うことができ、リスクの影響範囲が拡大する傾向がある。そのため本書では両者を明確に分類するのではなく、AI が備える機能の範囲やリスクの性質に応じて整理する。

■AI モデルの提供形態による分類と本書の対象範囲

AI モデルは、2026 年 6 月現在、提供形態に応じて大きく 3 つに分類できる。本書ではこれを以下のように整理した。

- **オープンモデル**：モデルの重みが公開されており、ローカル環境への展開が可能なモデル（例：Llama, Gemma）
- **商用利用可能モデル**：API やサービスを通じて広く一般に提供されるモデル（例：Gemini, Claude Sonnet）
- **カバードモデル（Covered Model）**：能力の高さゆえに特別な安全管理措置等が適用され、限定された組織のみアクセス可能なモデル[6]（例：Claude Mythos）

高度な能力を持つカバードモデルは安全管理上の理由等によりサービスが停止される場合もあり[7]、また現時点では利用できる読者が限られることを踏まえ、本書ではカバードモデルを対象外とする。

1.4 想定読者と活用イメージ

総務省・経済産業省が発行する「AI 事業者ガイドライン」[8]における AI 利用者を『AI 導入者』『AI 管理者』『セキュリティ担当者』『AI 利用者』と役割ごとに分類し、そのうち『AI 導入者』『AI 管理者』『セキュリティ担当者』を想定読者としている。また、本書が対象とする AI 利活用の範囲には、以下のいずれのケースも含まれる。

- AI システムを自社で開発する場合と、AI 提供者から AI システムを購入・利用する場合の双方
- 自社の業務効率化を目的とした AI 導入だけでなく、自社の製品・サービスの一機能として AI を

[5] GPTs とは、特定の目的に合わせて指示、知識、利用可能な機能などを設定した、ChatGPT 上のカスタム AI アシスタントを作成・共有・利用できる仕組みのこと。

[6] Anthropic, Covered Models, https://support.claude.com/en/articles/15425695-covered-models

[7] Anthropic, Statement on the US government directive to suspend access to Fable 5 and Mythos 5, https://www.anthropic.com/news/fable-mythos-access, 2026 年 6 月 12 日、米国政府の輸出規制指令により Claude Fable 5 および Mythos 5 へのすべてのアクセスが停止された事例

[8] 総務省・経済産業省, AI 事業者ガイドライン（第 1.2 版）
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20260331_1.pdf

組み込む場合

図 1-1　一般的な AI 活用の流れにおける主体の対応と本書の対象読者

※ AI 事業者ガイドライン 第 1.2 版（総務省・経済産業省）図 3 を元に作成

表 1-2　本書における活用主体の定義

AI 事業者ガイドラインによる定義	**本書での定義**
AI 開発者(AI Developer) AI システムを開発する事業者(AI を研究開発する事業者を含む)	**AI 開発者** （AI 事業者ガイドラインによる定義と同じ）
AI 提供者(AI Provider) AI システムをアプリケーション、製品、既存のシステム、ビジネスプロセス等に組み込んだサービスとして AI 利用者、場合によっては業務外利用者に提供する事業者	**AI 提供者** （AI 事業者ガイドラインによる定義と同じ）
AI 利用者(AI Business User) 事業活動において、AI システムまたは AI サービスを利用する事業者	**AI 導入者** ・AI システムの開発を AI 提供者に発注する企業の担当者 ・AI 提供者が運用する AI システムを自社に導入する企業の担当者
	AI 管理者 AI 導入者が導入した AI システムを管理し、適切に運用する企業の担当者
	セキュリティ担当者 企業における AI の全社ガバナンス整備、AI システムの導入・運用におけるセキュリティリスクの評価、インシデント対応を担う担当者
	AI 利用者 導入した AI システムを利用する者

1.5 本書の使い方

本書は通読することで全体像を把握できるが、目的に応じて必要な章を参照することも可能である。以下に各章の概要と活用による効果を示す。

表 1-3 各章の主旨および活用による効果

章	タイトル	主旨	活用による効果
2	**全社 AI ガバナンス**	組織全体で AI を統制するために必要な対応事項を示す。	AI 活用を全社的に推進するために、必要な施策を理解できる。
3	**AI システムのライフサイクル**	AI の導入から運用、廃棄全体を通して、フェーズの説明と、必要なセキュリティ対応事項を示す。	AI システムの企画から廃棄までを通して、フェーズごとに実施すべき取組みを理解できる。
4	**AI システムによる発生被害**	生成 AI および AI エージェントが引き起こす可能性のある被害を示す。	生成 AI や AI エージェントの利用によって、どのような業務被害が起こる可能性があるかを理解できる。
5	**AI システムの固有リスク評価**	発生被害ごとに、AI システムの固有リスクを測るための評価観点と評価基準例を示す。	導入および運用する AI システムのリスクがどの程度かを評価することができる。
6	**対策の効果と残存リスク**	発生被害ごとに対策を示し、その効果と残存リスクを示す。	発生被害に対してどのような対策があり、対策後にどのようなリスクが残るかを理解できる。

本書は、第 2 章から順に読み進めることによって、企業における AI ガバナンスの全体像から、個別の AI システムのリスク評価および対策を体系的に理解できる構成となっている。ただし、読者の役割や関心に応じて、表 1-4 のような読み方も可能である。

表 1-4 想定読者ごとの読み方例

想定読者	読むべき章
AI 導入者	3 章、4 章、5 章、6 章（特に技術要件）を中心に、AI のリスクや AI システムの導入時に実施すべき事項や対策を理解できる。
AI 管理者	3 章、5 章、6 章（特に運用対策、人的統制）を中心に、AI システムを管理・運用する時に考慮すべき事項や対策を理解できる。
セキュリティ担当者	2 章、3 章、4 章、5 章を中心に、AI 活用を全社的に推進するために必要な施策や、AI のリスクや評価方法、AI システムのライフサイクルを理解できる。

■NIST AI RMF との対応関係

なお、本書の構成は、国際的に広く参照されている AI リスク管理の枠組みである NIST AI リスクマネ

ジメントフレームワーク[9]（以下、「NIST AI RMF：NIST AI Risk Management Framework」という。）と整合的に設計されている。表 1-5 にその対応関係を示す。

表 1-5　NIST AI RMF と本書の対応

NIST AI RMF の機能	**本書の対応章**
GOVERN	第 2 章・第 3 章
MAP	第 4 章・第 5 章
MEASURE	第 5 章
MANAGE	第 6 章

第 2 章および第 3 章は AI ガバナンスの基盤を整備する章であり、第 4～6 章における個別のリスク評価・対策を支える位置付けにある。この構造は、NIST AI RMF において GOVERN 機能が MAP・MEASURE・MANAGE を支える構造と整合的である。NIST AI RMF のプレイブックには推奨アクションが詳細に記載されており、本書を補完する参考資料として有用である。本書ではプロジェクトメンバーが議論を通じて重要と判断した観点に絞って説明しているため、より詳細な実施事項については NIST AI RMF プレイブックも参照されたい。

本書は、AI の導入から運用・廃棄に至るライフサイクル全体を通じて、リスクの把握と対策の判断に必要な情報を提供するものである。各企業の業種・規模・システム構成・リスク許容度によって、適切な対策の選択は異なる。本書を足がかりとして、自社の状況に即した AI 利用方針の策定と運用体制の構築に活用されたい。

[9] NIST, AI Risk Management Framework, https://www.nist.gov/itl/ai-risk-management-framework
AISI, 米国 NIST AI リスクマネジメントフレームワーク（RMF）の日本語翻訳版,
https://aisi.go.jp/output/output_information/240704/

1.6 免責事項

本書は、企業が生成 AI および AI エージェントを導入・運用する際のセキュリティリスクに関する情報提供を目的として作成されたものであり、法的な助言を構成するものではない。また、本書の内容は予告なしに変更される場合がある。

本書に記載された内容は、執筆時点（2026 年 6 月）で入手可能な情報に基づいており、AI 技術およびそれを取り巻く脅威・規制の動向は急速に変化する。そのため、本書の内容が将来にわたって正確かつ完全であることを保証するものではない。実際の AI 導入・運用にあたっては、最新の技術動向、法令・ガイドラインの改正状況を確認のうえ、必要に応じて専門家の助言を得ることを推奨する。

本書に記載されたリスク評価や対策は、あくまで一般的な指針であり、すべての企業・環境に適合することを保証するものではない。各企業の業種・規模・IT 環境・リスク許容度に応じた判断が必要であり、本書の利用、および記載内容に基づく対策の実施・不実施により生じたトラブルや損害について、本書の著者ならびに監修者は一切の責任を負わない。

本書で紹介した事例・攻撃手法は、リスクの理解促進を目的としたものであり、攻撃行為の助長を意図するものではない。

本書に記載の内容は、独立行政法人情報処理推進機構および産業サイバーセキュリティセンターの意見を代表するものではなく、著者の見解に基づいている。

本書の有効期限は、発行日から 2 年間とする。

第 2 章　全社 AI ガバナンス

本章では、企業における AI の安全な導入および利活用を実現するための全社ガバナンスの考え方と実務上の要点を整理する。本書の著者は、AI ガバナンスの要諦は、単に利用を制限することではなく、実態を把握し、技術進化への機敏な適応を両立させながら、ビジネス価値を最大化しつつリスクを許容範囲内に管理することにあると考えた。

本章は、企業のセキュリティ担当者が、AI ガバナンスに関する実施事項を検討・推進する際の参照情報となることを目指す。

2.1　全社 AI ガバナンスの考え方

AI の技術進化は極めて速く、利便性やセキュリティリスクが短期間でも大きく変化する。従来型の制定以降改定がなされない社内 IT 管理規程をそのまま運用するだけでは、新たに顕在化するリスクに十分対応できない。このため、全社 AI ガバナンスを検討するに当たっては、以下の 2 点を基本思想とした。

なお、本章（第 2 章）および第 3 章は AI ガバナンスの基盤を整備する章であり、第 4～6 章における個別のリスク評価・対策を支える位置付けにある。この構造は、NIST AI RMF において GOVERN 機能が MAP・MEASURE・MANAGE を支える構造と整合的である。

（1）**現状の把握**：組織が利用を推奨する AI システム（組織公認 AI）の利用状況、組織が利用を許可した AI システム（組織承認 AI）の利用状況、組織が利用を許可していない AI システムを従業員が利用する行為（シャドーAI）の実態を正確に把握する。

（2）**継続的な適応**：ガバナンス体制およびガイドラインは、一度の策定で完了とするのではなく、新規 AI モデルの登場や新たな攻撃手法の顕在化に応じて、継続的に見直し・更新する。この思想は、総務省・経済産業省「AI 事業者ガイドライン」が提示する「アジャイル・ガバナンス」の考え方と整合する。[10]

本章では、上記の基本思想に基づき、AI ガバナンスを以下の 3 つの観点で説明する。

- **2.2　AI ガバナンス体制の整備**：AI 責任者の設置、各部門の役割分担、連携体制など、AI 統制に必要な体制や役割を示す。
- **2.3　AI ガバナンスルールの整備**：全社 AI 利活用ガイドラインの策定、シャドーAI への対応方針など、組織として取り決める事項を示す。
- **2.4　AI ガバナンスの継続的運用**：AI 活用実態の継続的把握、ガイドラインの動的なアップデート、相談窓口の設置、セキュリティ教育など、整えた体制とルールの運用方法を示す。

なお、本章は全社レベルの AI ガバナンスを扱う章である。AI システム個別の導入・運用に関わるガバナンスは第 3 章の AI ライフサイクルにおいて、AI システム単位で発生しうる被害およびその評価軸・対策は第 4 章から第 6 章において、それぞれ詳述する。

---

[10] 総務省・経済産業省, AI 事業者ガイドライン（第 1.2 版）
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20260331_1.pdf

2.2 AI ガバナンス体制の整備

AI ガバナンスを機能させるための第一段階は、組織内における責任の所在を明確化し、部門横断で意思決定を行う体制を整備することである。実効性のある統制には、経営層の関与や、推進に必要な組織能力、権限と責任の整理が必要である。既存のガバナンス体制を最大限に活用しつつ、AI 特有の変化に対応する体制について述べる。

2.2.1 AI 責任者（CAIO）の設置

■責任者を置く意義

AI の利活用が複数部門に跨り、多様な業務領域に影響を及ぼす中で、全社的な意思決定を担う責任者が不在のままでは、各部門の独自判断によるリスクの偏在や、責任の所在が不明確のまま対応が後手に回るリスクが生じる恐れがある。これらのリスクを避け、全社的な AI 利活用のゴール設定・推進およびリスク管理を統括し、意思決定の最終権限を有する責任者（以下、「CAIO：Chief AI Officer」という。）を設置することが望ましい。なお、デジタル庁が定める「行政の進化と革新のための生成 AI の調達・利活用に係るガイドライン」（DS-920）においても、各府省庁に AI 統括責任者（CAIO）を設置することが規定されており、本書の CAIO の位置付けもこれと同様の役割を想定している。[11]

■CAIO に付与すべき権限の観点

CAIO が実効的に機能するために、CAIO や CAIO 直轄組織が保有すべき主な権限を示す。[12]

- **全社 AI 利活用戦略および投資計画の策定**：AI 関連の全社戦略を立案し、AI への全社投資について、承認を行う。
- **AI 利活用および統制施策の実行判断**：AI 利活用および統制施策実行について、承認を行う。また、深刻なセキュリティ脅威が確認された場合に、対象システムの一時停止や外部 AI システムの利用禁止等を命じる権限があると望ましい。インシデント発生時に迅速な意思決定が取れるよう、あらかじめ権限を明文化しておくことが重要である。
- **経営層への AI リスクの共有**：重大なインシデントの発生および重大インシデントの兆候、利活用や規制に関する世の中の動向、全社的な AI リスクについて、取締役会や経営会議を通して、経営層に定期的に報告を行う。AI リスクを経営に重大な影響を及ぼしうるリスクの一つとして位置づけ、既存のリスクマネジメントプロセスの中で継続的に把握・報告することが望ましい。
- **部門を横断した活動の主導**：各組織（事業部門・IT・法務・人事・広報等）と常時連携し、AI に関する情報を提供し、対応を求める。AI 利用部門は全社に渡るため、全部門と連携を取りながら

[11] デジタル庁,「行政の進化と革新のための生成 AI の調達・利活用に係るガイドライン」（DS-920）
https://www.digital.go.jp/assets/contents/node/information/field_ref_resources/decb64eb-f26e-41cb-8d37-f3dd173108b8/59054b35/20260612_resources_standard_guidelines_guideline_01.pdf

[12] PwC, CAIO 実態調査 2025
https://www.pwc.com/jp/ja/knowledge/thoughtleadership/2025/assets/pdf/caio-survey-2025.pdf

推進する必要があるのはもちろんのこと、必要に応じて人事戦略やデジタル戦略といった機能戦略にAIの観点を盛り込む必要がある。

これらの権限を全て新規に整備する必要はなく、既存のCISO（Chief Information Security Officer, 最高情報セキュリティ責任者）やCIO（Chief Information Officer, 最高情報責任者）が保有する執行権限・報告ラインを応用することで、追加の規程整備を最小化できる。新たに付与すべき権限と既存権限の応用可能な範囲を整理した上で、不足分のみ規程化することが現実的である。

その他に必要な役割や責任、CAIO（およびCAIO直轄組織）に求められる能力については、AIセーフティ・インスティテュート（AISI）事務局発行の「Chief AI Officerガイドブック」に詳細に書かれているので、参照されたい。[13]

コラム1　CAIO・CISO・CIOの役割分担

CAIO、CISO、CIOはいずれもAIに関わる重要な役職であり、責任範囲が一部重複する。そのため、各社における役割分担を文書化することが重要である。表2-1に担当領域と役割の一例を示す。実際の分担は各社の組織構造に応じて調整する。

表2-1　CAIO、CISO、CIOの担当領域と役割

役職	主な担当領域	AI関連での役割例
CAIO	AIによる価値創出とAIリスク対応方針の策定	・全社AI利活用戦略および投資計画の策定 ・AI利活用および統制施策実行判断 ・経営層へのAIリスクの共有 ・部門を横断した活動の主導
CISO	サイバーセキュリティリスクの統制	・AIシステムの脆弱性・脅威管理 ・AIに起因するインシデントの技術的対応 ・セキュリティ要件のAI調達への反映 ・ログ監視・アクセス制御
CIO	IT基盤の整備・運用	・組織公認AIの整備・運用管理 ・AI関連システムのインフラ選定 ・IT部門とAI推進部門の連携調整

AIガバナンスの文脈では、AIの出力品質・倫理・社会的影響を管理する観点（AIセーフティ）と、データ漏洩やAIシステムへの攻撃から守る観点（AIセキュリティ）は、担当チームが異なる場合がある。

[13] AIセーフティ・インスティテュート（AISI）事務局, Chief AI Officerガイドブック（Version 1.00）
https://aisi.go.jp/assets/pdf/260317_caio_guidebook_v1.00_ja.pdf

ただし、両者の境界は実務上曖昧であり、例えばプロンプトインジェクション[14]対策は不適切な出力の抑止と悪意ある攻撃への対処の両面を持つ。そのため、体制上の役割分担を明確にしつつも、AI セーフティと AI セキュリティを統合して対処する連携体制を構築することが重要である。

役職者レベルの兼務（例：CISO が CAIO を兼務）は責任の所在を明確化する上で有効であるが、それだけでは実務レベルの連携が十分に機能しない場合がある。AI 推進担当とセキュリティ担当のメンバーが互いに兼務する体制を組織レベルで構築できると、役職者の判断が現場に届きやすくなり、ガバナンスの実効性が高まると考える。なお、CAIO の兼務の形態は多様であり、2026 年時点では一部の民間企業において CHRO（Chief Human Resources Officer, 最高人事責任者）との統合・兼任の動きも見られる。

2.2.2 部門を横断した体制の構築

■各部門へ AI 推進担当者を設置

各事業部門に、当該部門の業務プロセスと AI の接点を理解する推進担当者を配置することが望ましい。推進担当者は、自部門における AI 活用の推進およびリスク管理を主導する。既に IT 推進担当者等の役割が各部門に存在する場合は、その仕組みを応用することで設置が比較的容易になると考える。

推進担当者に期待される主な役割は、自部門の AI 利用状況の把握と報告、新規 AI 導入時の相談窓口との連携、自部門従業員へのガイドライン周知・教育、インシデント発生時の初動対応等が挙げられる。

■AI 連絡会議の設置

経営層の意思決定を各部門に反映させ、現場の課題を経営層に共有するため、CAIO 直轄組織と各部門の推進担当者が参加する会議体（以下、「AI 連絡会議」という。）を定期的に開催する。

AI 連絡会議で扱う主な議題の例として、AI ガイドライン改定の説明、教育プログラムの実施状況、シャドーAI 検出状況といった利用実態モニタリング結果、AI 活用事例紹介、発生インシデントおよび改善策等が挙げられる。

2.3 AI ガバナンスルールの整備

体制が整ったら、次に組織として取り決めるべき事項を明確にする必要がある。本節では、全社で共有すべきルールを 2 つの観点で述べる。

2.3.1 全社 AI 利活用ガイドラインの策定

全従業員が参照すべき利用基準として、全社 AI 利活用ガイドラインを策定する。システム単位の対策（第 3 章以降）とは異なり、本ガイドラインは組織として守るべきルールを定める。網羅的なガイドラインを

[14] プロンプトインジェクションとは、AI への入力文（プロンプト）に悪意ある指示を埋め込み、AI の動作を意図せぬ方向に誘導する攻撃手法のこと。

一度に整備しようとすると策定に時間がかかるため、最低限の内容から始め、継続的に更新していく運用が現実的である。更新の契機については 2.4.2 節を参照されたい。ここで、全社 AI 利活用ガイドラインに含めるべき観点例を表 2-2 に示す。

表 2-2 ガイドラインに記載すべき事項

観点	記載内容の例
適用範囲・対象者	・対象組織（例：国内や海外のグループ会社を対象とするか。） ・対象者（例：役員・正社員・契約社員・派遣社員・委託先の扱い） ・対象となる業務の範囲
利用可能な AI	・組織承認 AI システムの一覧 ・AI システム利用希望時の申請プロセスと審査基準
データ入力基準	・入力可能な情報の分類（例：公開情報、社内情報、機密情報、個人情報等）および入力禁止情報の定義 ・判断に迷う場合の相談窓口
出力結果の取扱い	・人間による出力の事実確認および第三者の権利侵害の確認 ・ハルシネーション（事実と異なるもっともらしい情報を生成すること）を前提とした運用設計の義務化
禁止事項	・組織が未許可の AI システムの利用（シャドーAI） ・利用を禁止する用途・サービス・製品 ・不適切な利用が判明した際の対応手順や利用停止措置の明記
インシデント時の対応	・情報漏洩・誤情報拡散・著作権侵害等を発見した場合の連絡先・連絡フロー・初動対応手順 ・AI インシデントの定義と重大度の分類基準
責任分界	・CAIO・CAIO 直轄組織・推進担当者・その他関連部門における役割の定義および責任の整理

留意点として、ルールは業務実態と乖離しないことが重要である。過度に厳格なルールは AI 利用者の回避（シャドーAI の増加等）を招き、かえってリスクを高める可能性がある。禁止一辺倒ではなく代替手段とセットでルールを策定することが望ましい。

また、ここではガイドラインを紹介したが、他にも AI 基本方針や AI 倫理指針、AI ポリシーを作成し、組織の透明性向上および対外的な信頼確保を目的に、社外に公表する企業も存在する。（特に、一般の顧客向けサービスに AI を利用する場合や、AI 適正利用の説明責任が問われる場合など）

2.3.2 シャドーAI への対応方針

シャドーAI の発生は多くの場合、悪意ではなく業務上のニーズから生じるため、禁止を先行させるだけでは申告されない利用が増え、実態把握がさらに困難になるリスクがある。

シャドーAI への対応は、場当たり的な個別対処ではなく、組織として方針を定めた上で実施することが重要である。方針の策定にあたっては、まず自組織におけるシャドーAI の実態を把握し、その状況に応じて AI システムを「利用禁止とするか」「条件付きで許可するか」といった方向性について、CAIO を中心となって合意する必要がある。

以下に具体的な統制方法を紹介する。

（1）組織公認 AI の提供

シャドーAI が発生する大きな要因として、「業務に使いたいのに公式な手段がない」状況が考えられる。IT 部門が安全性を検証した組織公認 AI を全社（または情報漏洩リスクの高い部署から優先的に）提供することが、シャドーAI 対策の有効な前提条件の一つであると考える。組織公認 AI の利便性が現場のニーズを満たせれば、シャドーAI の動機自体を低減できる。

（2）ルールの明文化と教育

組織承認 AI システムの一覧とそれ以外のサービスの利用禁止を全社 AI 利活用ガイドライン等に明文化し、従業員に周知する。また、新しい AI システムの利用を希望する場合の審査プロセスを整備し、現場の要望を吸い上げる仕組みを設けることが重要である。なお、審査の観点や却下になりやすい条件を事前に明示することで、申請者が事前に自己評価しやすく、セキュリティ部門や法務部門の承認業務の負荷軽減、並びに迅速な承認が可能になる。

（3）技術的な検知・遮断

クラウドサービスや Web サービスへのアクセスを監視・制御するセキュリティ製品（CASB[15]：Cloud Access Security Broker、SWG[16]：Secure Web Gateway 等）を導入し、未許可 AI システムへの通信を検知・遮断することも有効である。近年では、AI サービスへのデータ送信に特化した検知・制御機能を持つ AI DLP 製品も登場しており、従来のセキュリティ製品では識別が困難だった AI システム固有の通信パターンの検知や AI サービスへの機密情報の入力制御が可能になりつつある。ただし、これらの製品の導入にはネットワーク変更や業務影響調査の実施が必要なため、費用や工数が増大する可能性がある。そのため、まずは明らかに危険なサービスの遮断に絞ったスモールスタートが現実的である。また、すべての AI システムを技術的に完全に遮断することは難しく、（1）や（2）との組み合わせで対応するのが望ましい。

なお、企業ネットワークを経由しない端末からの AI システム利用は技術的な制御が困難であり、最終的には従業員の意識とルール順守に依存する部分が残る。シャドーAI 対策は一度整備すれば終わりで

---

[15] CASB とは、サービス利用者とクラウドサービスの間に位置するセキュリティ制御ポイントにて、クラウド利用の可視化、データ保護、脅威防御、コンプライアンス適用などの機能を提供する製品のこと。

[16] SWG とは、インターネットへのアクセスを中継・監視するプロキシ型セキュリティ製品のこと。URL フィルタリング、マルウェア検知、SSL/TLS 復号検査などの機能を提供する。

はなく、新たな AI システムの登場や従業員の利用実態の変化に応じて、組織公認 AI の利便性向上や組織承認 AI の拡充、ルールの見直しを継続的に行うことが求められる。

2.4 AI ガバナンスの継続的運用

体制とルールを整備した後は、それらを形骸化させずに継続的に機能させるための運用が重要となる。本節では、5 つの観点で AI ガバナンスを継続的に運用するための要旨を述べる。

2.4.1 AI 活用実態の継続的把握

2.1 節で述べた「現状把握」は、一度きりの初期調査ではなく、最低でも年 1 回を目安に定期的に繰り返す活動である。AI の普及速度や開発状況、シャドーAI の実態、業務利用の深度変化等は常に変化するため、実態を継続的に把握し、統制の精度を保つことが求められる。また、主要な AI モデル・サービスの大型アップデート、業界・社内での重大インシデント発生、ガイドライン改定のタイミングでも追加実施することが望ましい。ここで、AI 活用実態を理解するために把握すべき観点と項目例を表 2-3 に示す。

表 2-3 AI 活用実態理解のために把握すべき事項

観点	項目例
利用している AI システム	・製品名 ・AI 提供事業者 ・契約形態（例：法人契約、個人契約、無料利用）
利用目的と範囲	・AI を利用する業務または AI 機能を提供しているサービス ・利用者の範囲 ・入力データの種別（例：公開情報、社内情報、機密情報、個人情報） ・出力結果の利用先（例：社内のみ、社外提供、顧客向けサービス組込）

2.4.2 ガイドラインの動的なアップデート

2.3.1 節で策定したガイドラインは、策定後も継続的に更新し続けることが前提である。AI 技術の進化速度を踏まえると、最低でも年 1 回程度の定期改定を基本とすることが望ましい。一方、ガイドラインの改定には承認プロセスが伴うため、年 1 回程度の定期的な改定も困難な場合が想定される。ただし、放置すれば、技術進化や脅威動向と乖離したガイドラインが形骸化するリスクがある。

このため、年 1 回でなくとも、以下のようなタイミングが到来した場合に改定を行うことを予め定めておくことが効果的である。

- 新たな攻撃手法・脆弱性の顕在化
- 関連法令・公的ガイドラインの改定
- 社内または業界における重大インシデントの発生

なお、改定の頻度を抑えたい場合は、普遍的な考え方をガイドライン本体に記載し、詳細や具体例は

補足資料として別途整備する二層構造も有効である。

また、ガイドラインの更新にあたっては、最新の脅威動向や AI セキュリティに関する知見を継続的に収集することが重要である。参考となる主な情報源を以下に示す。

- OWASP Top10 for LLM Applications[17]
- OWASP Top10 for Agentic Applications[18]
- MITRE ATLAS[19]
- AI セーフティ・インスティテュート（AISI）からの情報[20]

2.4.3 各部門からの相談窓口設置および伴走支援

各部門による独自の AI 導入・運用を抑止しつつ、安全で効果的な活用を推進するためには、各部門からの問合せを受ける相談窓口を設置し、AI 推進担当やセキュリティ担当による伴走支援を提供することが望ましい。窓口を設置するだけでは、問題に気づいていない部門に支援が届かないため、受動的な窓口対応に加え、AI 推進担当やセキュリティ担当から各部門へ定期的に働きかける能動的な接続（例：定期ヒアリング）が重要である。ここで、相談窓口に期待される役割について、表 2-4 に示す。

表 2-4 相談窓口に期待される役割

役割	項目例
AI システム導入（改修）相談	・AI システム導入（改修）の妥当性・リスク評価・AI 提供者の選定・データ取扱い等に関する相談受付 ・AI システム導入（改修）におけるフェーズ移行時の成果物レビュー ・高リスク AI システムにおける開発伴走支援 ・AI システムの導入・運用におけるリスク評価支援
インシデント対応支援	・AI 起因のインシデントについて、調査・再発防止策の検討支援 ・他部門へ注意喚起の発信
情報発信	・最新の脅威・脆弱性情報 ・AI 利活用ガイドラインの改定 ・社会や業界、他部門の成功事例の紹介
ナレッジ蓄積	・相談および支援の結果から得られた知見を FAQ やベストプラクティス集などの形式で蓄積し、後続の相談者が参照できるようにする

[17] OWASP Foundation, OWASP Top 10 for Large Language Model Applications, https://owasp.org/www-project-top-10-for-large-language-model-applications/

[18] OWASP Foundation, Gen AI Security Project, OWASP Top 10 for Agentic Applications for 2026, https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

[19] The MITRE Corporation, MITRE ATLAS, https://atlas.mitre.org/

[20] AISI, 活動, https://aisi.go.jp/activity/

2.4.4 AI セキュリティ教育の実施

AI を利用する全ての従業員に対して、AI リスクと正しい利用方法の教育を定期的に実施する。教育に含めるべき主な項目を以下に示す。

- AI 利活用ガイドラインの内容
- AI が引き起こす発生被害とその事例（第 4 章）
- ハルシネーションを前提とした AI システムからの出力の確認の重要性
- （組織公認 AI がある場合）組織公認 AI を使うべき理由

AI の技術や脅威は急速に変化しており、半年前の知識が陳腐化していることもあるため、これらの教育は、一度の研修では終わらせず、定期的に教育を実施することが重要である。また、禁止事項の押し付けだけでなく、「組織公認 AI にはこのような安全策があり、安心して使うことができる」というポジティブなメッセージを伝えることで、ルールの順守率を高めることができる。なお、経営層と AI 推進担当者については、上記に加えて、表 2-5 に示すような個別にカリキュラムを設計することを推奨する。

表 2-5 教育対象者と学習内容

対象	学習内容の例
経営層	・AI による事業機会と経営リスクの全体像 ・AI に関する主要な法令・規制動向 ・他社のインシデント事例
AI 推進担当者	・AI ライフサイクルの各フェーズの実施事項（第 3 章） ・固有リスク評価の実施手順（第 5 章） ・AI リスクへの対策方法（第 6 章）

2.4.5 AI のリスクに対応した保険の加入検討

生成 AI のリスクは、本章に記載のガバナンス体制やルールの整備、および第 6 章に示す対策によって低減できるが、完全に排除することはできない。万が一のリスク顕在化に備える手段の一つとして、保険の活用が考えられる。

まず、既にサイバー保険に加入している場合は、その補償範囲が AI 活用に伴うリスクをカバーしているか確認する。サイバー保険は一般に、不正アクセスや情報漏洩、ランサムウェア被害などをカバーするが、AI の出力に起因する知的財産侵害やハルシネーションによる誤った業務判断から生じる損害については、補償対象外となる場合がある。自組織の AI 活用範囲が拡大するにつれ、既存の補償内容が実態に即しているかを定期的に見直すことが望ましい。

なお、保険はあくまでリスクの「移転」手段であり、リスクそのものを低減するものではない。ガバナンス体制の整備や技術的対策を前提とした上で、残存リスクへの備えとして位置付けることが重要である。

第 3 章　AI システムのライフサイクル

本章では、AI システムを新規に企画し、運用、廃棄に至るまでの基本的な流れを示す。AI システムは、従来の IT システムと同様に要件定義、設計、実装、試験および運用管理を必要とするが、入力情報の保存・再利用、出力の不確実性、プロンプトインジェクション等の AI システム固有のリスクを伴う。そのため、各フェーズで実施すべき事項、判断基準および承認の観点を明確にする必要がある。

本書では、AI システムのライフサイクルを実務上の工程管理に合わせ、「企画→調達→技術検証・設計→運用前準備→運用→廃棄」（図 3-1「AI システムのライフサイクル」）と定義し、各フェーズでの AI 導入者（業務部門等）、AI 管理者（運用部門等）およびセキュリティ部門等の支援部門の関与範囲を 表 3-1「関係者別役割分担表」に定める。なお、法務部門および調達部門等の関係者は、対象とする AI システムの性質に応じて各フェーズで関与する。

図 3-1　AI システムのライフサイクル

フェーズ移行時には、成果物を作成する部門だけでなく、関係するステークホルダーが協議し、リスクが許容範囲内であることおよび残存リスクの取扱いを文書化した上で次工程に進む。特に、外部顧客向けサービス、機密情報または個人情報を扱う用途、自動判断または外部システム操作といった用途では、専門的な知識が必要となるため、AI 管理者やセキュリティ部門を含めて各フェーズでの成果物承認を必須とすることが望ましい。

表 3-1 関係者別役割分担表

フェーズ	主担当とその役割	支援部門とその役割
1.企画	**・AI 導入者**： ビジネス要件の明確化、 固有リスク評価の実施、 セキュリティ要件と対策検討	**・AI 管理者**： 企画内容の確認、フェーズ成果物の承認 **・セキュリティ部門**： 企画内容の確認、固有リスク評価の支援、 フェーズ成果物の承認 **・情報システム部門**： 企画内容の確認、固有リスク評価の支援 **・法務部門**： 企画内容の確認
2.調達	**・AI 導入者**： AI 提供者の評価、 AI 開発者およびサプライチェーンの評価、 契約へのセキュリティ要件の反映	**・AI 管理者**： 契約内容の確認、フェーズ成果物の承認 **・セキュリティ部門**： 評価の支援、契約内容の確認、 フェーズ成果物の承認 **・法務部門**： 契約内容の確認 **・調達部門**： 契約内容の確認、契約締結の支援
3.技術検証・設計	**・AI 導入者**： セキュリティ要件への適合確認、 運用設計およびシステム設定の実施、 システム管理体制の構築 **・AI 管理者**： 運用設計およびシステム設定の実施、 システム管理体制の構築、 フェーズ成果物の承認	**・セキュリティ部門**： 検証・設計等の支援、 フェーズ成果物の承認
4.運用前準備	**・AI 導入者**： 試験の実施および結果の評価 **・AI 管理者**： 試験の実施および結果の評価、 運用手順の最終確認	**・セキュリティ部門**： 評価結果への助言、 フェーズ成果物の承認
5.運用	**・AI 管理者**： 定期的な固有リスク評価、 定期的な対策の実効性評価、 変更管理の記録、 脅威・脆弱性情報の継続的収集と評価、 AI 開発者・AI 提供者の定期的な評価	**・セキュリティ部門**： 評価の支援・助言 **・情報システム部門**： 固有リスク評価の支援
6.廃棄	**・AI 管理者**： データおよび資産の整理	**・セキュリティ部門**： 廃棄の支援・助言

3.1 企画

企画は、AI システムの提供目的、適用業務または機能、利用者、AI が判断または処理を担う範囲、および AI を利用しない用途を定める段階である。従来の IT システムでは、同一の入力に対して一定の結果が返されることを前提として設計される。一方、AI システムでは、出力内容に不確実性があり、同じ入力であっても常に同一または正確な回答が得られるとは限らない。そのため、企画段階から、利用目的の妥当性、利用制限、誤回答・不適切出力等が発生した場合の影響、人による確認範囲、提供上の注意事項や免責事項、禁止事項を定める必要がある。

企画フェーズでは、AI 導入者が中心となって、AI システムのビジネス上の目的、利用範囲および期待効果を整理し、AI 管理者やセキュリティ部門、情報システム部門、法務部門が、統制およびリスクの観点から企画内容の妥当性を確認することが重要である。特に、AI 管理者は、導入後の管理および適切な運用を担う立場から、将来の運用に必要となる事項を企画段階から確認し、企画内容に反映することが望ましい。

なお、企画フェーズで定めた内容は、調達時の RFP（Request for Proposal：提案依頼書）[21]、AI 提供者の評価、契約条件、技術検証、運用設計等の前提となる。企画フェーズの終了時には、AI 導入者だけでなく AI 管理者やセキュリティ部門が、ビジネス要件、固有リスク評価、セキュリティ要求事項および残存リスクを確認し、調達フェーズに進める状態であることを承認する。

本節では、企画フェーズで実施すべき 3 つの事項を述べる。

3.1.1 ビジネス要件の明確化

ビジネス要件の明確化では、AI システムについて導入目的、対象利用者、対象機能、利用形態および AI が処理する範囲を明確にする。対象利用者については、外部顧客向けか社内向けか、特定の利用者向けか不特定多数向けかによって、求められる管理水準が異なる。一般に、外部顧客向け・不特定多数向けのほうが高い管理水準が求められる。利用形態においては、AI の出力を情報収集といった一時的な参考情報に留めるか、特定の業務プロセスに組み込んで利用するのか、あるいは外部システムの操作まで自律的に行わせるのか等を整理する。AI が処理する範囲においては、自動化された応答・判断・実行、外部公開、機密情報の取扱いを伴う場合は、リスクが高くなる。そのため、これらに該当する AI システムは、高リスク用途として AI 管理者およびセキュリティ部門へ事前に相談することが望ましい。

この段階の成果物として、AI 活用方針、対象機能一覧、対象利用者一覧、業務フロー、AI の関与範囲、人による確認範囲およびスコープ外とする用途を整理する。目的および対象が曖昧な場合、不要な機能拡張または高リスク用途への拡大が生じ、後工程で統制不能となる恐れがある。

3.1.2 固有リスク評価の実施

[21] RFP とは、調達先の候補に対して、要件・条件・評価基準を提示し、提案書の提出を求めるための文書。

ビジネス要件に基づいて、導入する AI システムの固有リスクを評価する。

AI 固有リスク評価を実施しないまま調達または開発に進むと、導入後に想定外のリスクが顕在化しやすい。AI システムによる発生被害の詳細は第 4 章を、固有リスクの評価方法は第 5 章を参照されたい。

3.1.3 セキュリティ要件と対策の検討

固有リスク評価の結果を踏まえ、AI 提供者に提示するセキュリティ要求事項を整理する。セキュリティ要件として検討が必要な事項の例を以下に示す。

- アクセス・権限管理：アクセス制御、認証、権限管理、外部送信制限
- データ保護：暗号化、保存期間、削除条件、サービス終了時のデータ返却または削除
- ログ・監査：ログ取得、監査証跡の保存、監査対応
- AI 設定管理：プロンプトおよび設定値の管理、ガードレール[22]設定、モデル更新時の再評価
- 委託・契約管理：再委託条件、障害時報告、インシデント通知、SLA

なお、技術的対策だけでは対応できない場合は、運用対策およびガバナンス上の対策もあわせて検討する必要がある。詳細な対策方法については、第 6 章を参照されたい。

3.2 調達

調達は、AI システムの提供に必要な AI 提供者を選定し、本番利用に必要な要件への適合を確認する段階である。AI の性能または費用だけでなく、AI 提供者のセキュリティ管理の実施状況、財務健全性および事業継続性等も評価の観点となる。また、机上評価だけで判断できない事項は、試行利用等で事前に確認することが望ましい。本節では、調達フェーズで実施すべき 3 つの事項を述べる。

3.2.1 AI 提供者の評価

AI 提供者については、AI システムの利用目的、利用範囲および業務上の重要度を踏まえて、セキュリティ対策状況等を評価する。外部顧客向けサービスまたは業務上重要なシステムで利用する場合は、セキュリティ管理体制だけでなく、AI 特有のリスク管理、サービス継続性、財務健全性およびサポート体制も評価対象とする必要がある。評価観点の例を以下に示す。

- セキュリティ管理：セキュリティ管理体制、インシデント対応体制、ログ取得およびログ提供範囲
- データ管理：入力データ、出力データ、ログおよび学習データの取扱い
- AI モデル：提供可能な AI モデルの種類、提供形態、モデル更新方針、性能・安全性評価の報告状況
- 委託・監査：再委託先の管理、外部監査の実施状況、監査報告書の提供可否
- 運用・障害対応：障害発生状況、SLA の有無、サポート時間および問い合わせ対応体制
- 財務健全性：主要事業の収益性、事業撤退リスク、資金調達状況

[22] ガードレールとは、AI システムへの入力や AI システムからの出力を監視し、有害・不適切な出力の抑制、禁止用途への応答拒否といった制御をするための機能のこと。

評価結果は、調達可否の判断にとどまらず、契約条件、許容できる残存リスクの条件、追加対策および運用時の監視項目等の検討に反映する。

3.2.2 AI 開発者およびサプライチェーンの評価

AI システムでは、直接契約する AI 提供者だけでなく、基盤モデル[23]、外部 API、ライブラリおよびプラグイン等の依存関係を確認する。AI 提供者が第三者の AI 開発者に依存している場合、その依存先の変更、停止、脆弱性、利用条件変更等がサービスに影響する可能性がある。

AI 開発者については、3.2.1 節に準じてセキュリティ対策状況等を確認するほか、モデルの提供条件、学習データおよび利用データの取扱い、入力データの学習利用、モデル更新方針、財務状況および事業継続性も確認する。

これらのサプライチェーン評価の結果は、契約条件、設計、監視、代替可能性および廃棄時のデータ整理等の検討に反映する。

3.2.3 契約へのセキュリティ要件の反映

企画フェーズで整理したセキュリティ要求事項について、AI 提供者との間で合意し、契約に反映する。

以下の観点から必要事項を契約に明記する。説明資料や口頭説明のみでは、インシデント発生時、仕様変更時またはサービス終了時に必要な義務履行を確保できない恐れがある。

- 責任・協力：責任分界、インシデント時の協力、損害賠償、責任制限、損害発生時の対応
- 障害・脆弱性対応：障害時対応、モデル更新または仕様変更時の通知、脆弱性の継続監視、パッチ適用、脆弱性対応状況等の定期報告
- データ管理：データの利用目的、保存期間、削除義務、サービス終了時の返却または削除条件、ログ提供
- 委託・監査：再委託条件、監査可否、SLA、法令遵守、秘密保持

契約締結前には、AI 導入者、AI 管理者、セキュリティ部門、調達部門および法務部門が、企画段階で定めたセキュリティ要求事項が契約に反映されているかを確認する。契約で担保できない事項がある場合は、運用対策で補完できるかを判断し、補完できない場合はリスクの受容または採用の見送り等の判断を行う。

3.3 技術検証・設計

技術検証・設計は、調達した AI システムがビジネス要件およびセキュリティ要件を満たすかを確認し、運用開始に必要な設定、監視項目および管理体制を具体化する段階である。

従来の IT システムでは、要件どおりに機能が動作するか、性能および可用性を満たすかが主な確認対

[23] 基盤モデルとは、大量のデータで事前学習された汎用的な大規模 AI モデルのこと。様々なサービスやシステムの基盤として利用される。

象となる。AI システムでは、これに加え、出力品質、禁止用途への利用抑止の確認、プロンプトに入力されたデータの取扱い、外部連携時の権限制御、ログからの追跡可能性等についても確認する必要がある。本節では、技術検証・設計フェーズで実施すべき 3 つの事項を述べる。

3.3.1 セキュリティ要件への適合確認

企画フェーズで整理したセキュリティ要件について、調達した AI システムが適合しているかを検証する。要件を満たせない場合は、追加の技術的対策で補完するか、運用対策またはガバナンスで補完するかを検討する。具体的には、サービス側で入力データの保存を制御できない場合は、入力情報を制限する、匿名化する、対象業務を限定する、別サービスを採用する等の判断が必要となる。

セキュリティ要件における確認観点の例を以下に示す。

- データ・ログ管理：ログ取得、データ保存条件、削除条件、監査証跡
- 権限・接続管理：権限管理、外部接続制御、接続先の管理
- 安全対策：ガードレール、フィルタリング、運用監視の実現可否

技術検証の結果は、採用可否、リリース条件、残存リスク、追加対策、運用監視項目、契約条件の見直しおよび運用手順に反映する。検証結果が不十分なまま次フェーズに進むと、試験で発見した問題の原因が設計上の問題か運用上の問題かの切り分けが困難になる。

3.3.2 運用設計およびシステム設定の実施

AI 管理者は、AI 導入者と相談の上、AI システムの運用開始後に実際に監視および変更管理を行う主体であるため、企画フェーズで整理した運用対策を、実装可能な状態に具体化し、設定および検証を行う。例えば、契約で入力データの学習利用を禁止していた場合でも、保存または再利用を無効化する設定が適切に行われていなければ、契約上の禁止事項の実効性は確保されない。また、運用対策の有効性は、文書化だけでなく、設定画面、ログ、権限一覧、テスト結果、変更履歴等により確認できる必要がある。

整備対象の例を下記に示す。

- アクセス・接続管理：アクセス制御、認証設定、接続先制御、利用制限
- 秘密情報・ログ管理：秘密情報管理、ログ設定、保存設定
- AI 制御：フィルタリング設定、ガードレール設定、プロンプト管理、システムプロンプト[24]変更管理
- 運用手順：停止条件、エスカレーション手順、問い合わせ対応手順

3.3.3 システム管理体制の構築

AI 導入者は、AI 管理者と相談の上、運用開始後における AI システムに関する管理体制、責任分界点および意思決定者を明確にしておく必要がある。インシデント発生時には、被害拡大を防ぐための代

[24] システムプロンプトとは、AI モデルの動作方針・役割・制約をあらかじめ設定するための指示文のこと。利用者からの入力より前に処理され、AI の応答全体に影響する。

替運転やサービス停止をはじめ、封じ込め、原因分析、顧客影響確認、証跡保全、対外公表および再発防止等の判断が必要となるため、平時から対応体制を整備し、緊急時に確実に機能する状態を維持することが重要である。

また、BCP（事業継続計画）の観点から、システム停止時の代替運用手段、切替・復旧手順の整備および復旧目標と影響範囲の明確化についても整理することが望ましい。特に、外部顧客向けサービスまたは業務上重要なシステムについては、上記事項の整理を必ず実施すべきである。AI 開発者または AI 提供者の障害、仕様変更、提供停止によりサービスが影響を受ける可能性もあるため、外部依存を考慮した継続計画が必要である。

システム管理体制の構築結果は、運用手順書、連絡網、権限一覧、監視項目一覧、インシデント対応手順、BCP、委託先との連絡手順および判断基準として整理し、次フェーズで有効性を確認する。

3.4 運用前準備

運用前準備は、本番稼働前に、安全性、禁止用途への利用抑止、攻撃耐性、運用実現性およびインシデント対応力を確認する段階である。従来の IT システムに求められる機能、性能、可用性およびセキュリティの確認に加え、AI の出力品質、安全性および利用目的への適合性を確認する。

試験の実施主体は、AI 導入者に限らず、AI 提供者または外部専門家に委託して実施する場合もある。いずれの場合も、AI 導入者および AI 管理者は試験結果を受領するだけでなく、固有リスク評価で判定したリスクが許容水準まで軽減されているかを確認し、CAIO 等の承認者がリリース可否を判断する必要がある。本節では、運用前準備フェーズで実施すべき 2 つの事項を述べる。

3.4.1 試験の実施および結果の評価

AI システムに対する試験では、従来の IT システムに対する機能試験、結合試験、性能試験および受入試験等に加え、AI 特有のリスクを踏まえた安全性確認も行う。通常の利用シナリオだけでなく、誤用、悪用、プロンプトインジェクション等の誘導攻撃を想定したシナリオについても試験を実施することが望ましい。

特に、外部顧客向けサービス、自動化された応答・判断・実行、社外システムとの連携または機密情報の取扱いを伴うシステムの場合は、リスクの高さに応じて外部専門機関等によるペネトレーションテスト（侵入テスト）の実施を検討する。

試験結果を確認する際は、固有リスク評価の結果が変化していないか、および、固有リスク評価で判定したリスクの高さを 3.3 節までに実施した対策によって許容水準まで軽減できているかについても判断する。重大な問題が発見された場合は、設計・設定の変更、提供範囲の縮小、運用対処、契約条件の見直しまたはリリース延期を検討する。

3.4.2 運用手順の最終確認

AI システムの運用対策およびインシデント対応体制が、実際の運用場面で有効に機能するかを確認する。3.4.1 節の試験が主に AI システムの挙動を確認するものであるのに対し、ここでは、関係者が定め

られた手順に従い、異常の検知、判断、連絡、封じ込め、復旧および再発防止までの一連の対応を実行できるかを確認する。

確認結果をもとに、手順の不備、判断基準の不明確さ、連絡先の不足、ログ取得の不足、停止操作の難しさ、顧客影響確認の遅れ等を把握し、改善につなげる。確認された改善事項は、運用開始前に是正することが望ましい。運用開始前に是正できない事項については、リリース条件、暫定対策、対応期限、責任者および承認者を明確にして管理する。

3.5 運用

運用は、AI システムの提供開始後に、導入時点での対策およびリスク評価が継続して有効であることを確認し、必要に応じて見直す段階である。

従来の IT システム運用では、可用性、性能、障害監視および変更管理が中心となる。AI システムでは、正常稼働していても出力内容が不適切となる場合があり、利用者の使い方またはモデル更新によりリスクが変動するため、出力内容、利用実態およびコスト変動を含めた継続監視が必要である。

AI システムは環境変化が速いため、年 1 回程度の定期点検だけでなく、重要な変更、事故、脆弱性公表、モデル更新、利用範囲拡大等を契機とした臨時点検を行い、リリース条件または利用条件についても必要に応じて見直す。本節では、運用フェーズで実施すべき 5 つの事項を述べる。

3.5.1 定期的な固有リスク評価

運用開始後も、3.1.2 節で評価した AI システムの固有リスクが変動していないかを定期的に評価する必要がある。例えば、ビジネス要件の変更、対象利用者の拡大、当初想定と異なる利用方法、法令または業界ルールの変更により、導入時点のリスク評価が妥当でなくなる場合がある。

評価の結果、リスクが許容水準を超えると判断される場合は、対策の見直し、提供範囲の縮小または利用の停止を検討するとともに、必要に応じて契約の見直しも行う。

3.5.2 定期的な対策の実効性評価

AI システムでは、導入時点で有効であった対策も、利用状況またはモデル更新により効果が低下することがある。例えば、ガードレールの回避方法が共有される、AI 利用者が想定外の入力を行う、API 仕様変更によりログ項目が変わる等である。また、導入後に新たな対策手法が登場する場合もあるため、定期的な評価の機会に最新の対策を取り入れることも検討すべきである。対策の実効性評価では、設定が存在するかだけでなく、リスク低減に実際に寄与しているかを確認する。

実効性評価の観点の例を以下に示す。

- 監視・検知：ログ監視、異常検知、機密情報漏洩の検知、不適切出力および有害出力の検知（重要システムでは毎日、人が確認することが望ましい）
- 設定・権限：設定棚卸し、アクセス権レビュー、接続先確認、ガードレール設定
- 運用実績：インシデント対応記録、教育実施状況、例外承認の妥当性、是正状況

3.5.3 変更管理の記録

AI システムでは、モデル、プロンプト、設定および外部連携先の変更により、出力内容、情報の取扱い、外部システムへの動作または利用者への影響が変化する場合がある。そのため、変更内容に応じて、固有リスク評価、セキュリティ対策の有効性および運用手順への影響を確認し、必要な検証を実施する。

変更管理では、変更内容、影響確認結果、承認履歴、検証結果およびリリース結果を記録する。

特に、自動化された応答・判断・実行、外部システム連携、顧客向け表示、機密情報の取扱い、モデル差替えまたは外部 API の追加を伴う高リスクな変更については、影響範囲および残存リスクを確認した上で、承認者、承認日、リリース条件およびリリース後の確認結果を記録する。

変更後に問題が発生した場合に備え、変更前後の設定、検証結果、リリース判断の根拠および切戻し手順を確認できる状態にしておく。これにより、変更に伴う影響を追跡可能にし、障害発生時の原因特定、復旧対応および再発防止に活用する。

3.5.4 脅威・脆弱性情報の継続的収集と評価

利用中の AI システム並びにその構成要素については、導入後も継続的に関連情報を収集する。導入時点で安全性を確認していた場合であっても、脅威動向の変化、脆弱性の公表、サービス仕様の変更、モデル更新、提供条件の変更等により、利用中のリスクが増大する可能性がある。

収集対象には、AI システムの動作および安全性に影響を与える構成要素を含める。具体的には、従来のソフトウェア脆弱性情報に加え、AI 特有の脅威動向、障害情報、サービス終了情報、仕様変更、安全性評価情報およびベンダー更新情報等を継続的に確認する。また、3.2.3 節で合意した AI 提供者からの脆弱性対応状況の定期報告を活用し、自社による情報収集と組み合わせて確認することが効率的である。

収集した情報は保管にとどまらず、AI システムへの影響範囲を評価し、必要な対応に反映する。重大な脆弱性、深刻な障害、重要な仕様変更または提供条件の変更が確認された場合は、通常の定期評価を待たずに臨時の再評価を行う。再評価の結果、利用継続により許容できないリスクが生じると判断される場合は、担当部門は CAIO 等の責任者に報告の上、機能制限、一時停止、代替手段への切替等を検討する。

3.5.5 AI 開発者および AI 提供者の定期的な評価

運用開始後も、3.2.1 節および 3.2.2 節で実施した、AI 提供者、AI 開発者およびサプライチェーンのセキュリティ対策状況、財務健全性、事業継続性、SLA 達成状況、障害対応状況、脆弱性対応状況、提供条件の変更等について、定期的または重要な変更時に再評価する必要がある。

外部依存先の状況が悪化した場合や AI 提供者が意図せず変更された場合など、AI システムに直接の障害が発生していなくても、将来のサービス継続性または情報管理上のリスクが高まる。そのため、状況の深刻度に応じて、追加監査、契約見直し、代替先の検討、移行計画の作成および利用範囲の制限等を検討する。

3.6 廃棄

廃棄は、AI システムの廃止または AI システムの提供終了に際して、関連するデータ、認証情報、契約および証跡を整理する段階である。従来の IT システムでもサービス終了に伴う廃止手続は必要であるが、AI システムでは、入力データ、会話履歴および API キー等が特に残存しやすい点に留意して確認することが重要である。対応が漏れると、後日の情報漏洩、不正利用、誤課金または契約違反の原因となる。3.6.1 節で、廃棄フェーズで実施すべき事項を述べる。

3.6.1 データおよび資産の整理

廃棄時には、保存・削除・無効化・返却または移行の対象と承認者および実施者を決定し、実行証跡を残す。自社環境だけでなく、クラウド環境、AI 開発者および AI 提供者等の外部委託先、外部 API 提供者等に残存するデータの削除または返却条件を確認する。契約上、削除証明または作業完了報告を取得できる場合は、証跡として保全する。

整理対象の例を以下に示す。

- 利用データ：顧客データ、入力履歴、出力履歴、会話履歴
- ログ・学習関連：監査ログ、運用ログ、ベクトルデータベース[25]、評価用データ
- AI 設定・接続：プロンプト、設定情報、ガードレール設定、接続先設定
- 認証情報：アカウントおよび API キー・トークンの無効化、外部連携の認可設定の削除
- 文書・証跡：運用マニュアル、設計書、契約書、監査証跡、削除証跡

保存すべき記録については、法令、契約、監査、事故調査、顧客対応および社内規程に基づき保存期間を定める。削除すべき情報を不要に保存し続けることはリスクである一方、必要な証跡を削除すると、後日の説明責任を果たせなくなるため、保存と削除の基準、責任者および実施確認方法を明確にする。

以上が第 3 章の内容である。AI システムのライフサイクルでは、各フェーズを単独で捉えるのではなく、企画時のリスク評価、調達時の契約、技術検証時の対策実装、運用前の確認、運用中の監視および廃棄時の証跡管理を一連の管理プロセスとして扱い、各フェーズの判断結果を後工程に引き継ぎ、変更時に再確認する必要がある。これにより、導入時点の判断を一過性のものにせず、AI システムのライフサイクル全体を通じて説明責任、追跡可能性および統制の継続性を確保できる。

[25] ベクトルデータベースとは、テキストや画像等のデータを数値ベクトルに変換して格納するデータベースのこと。AI が類似情報を検索・参照するために利用される。

第 4 章　AI システムによる発生被害

4.1　発生被害の定義と一覧

本章では、AI システムがセキュリティインシデントや不適切な利用によって引き起こす可能性のある被害（リスク）を説明する。本書ではこれらを「発生被害」と定義する。

なお、図 4-1 で示す通り、第 5 章では、本章で説明する発生被害ごとに、影響度に基づきリスクの高さを評価するための観点を示す。第 6 章では各発生被害に対するリスクの高さに応じた対策を説明する。

また本章で定義する発生被害は現時点の技術水準や利用実態においてありうるものであり、今後の技術発展や環境変化に伴い変化する可能性がある点に留意されたい。

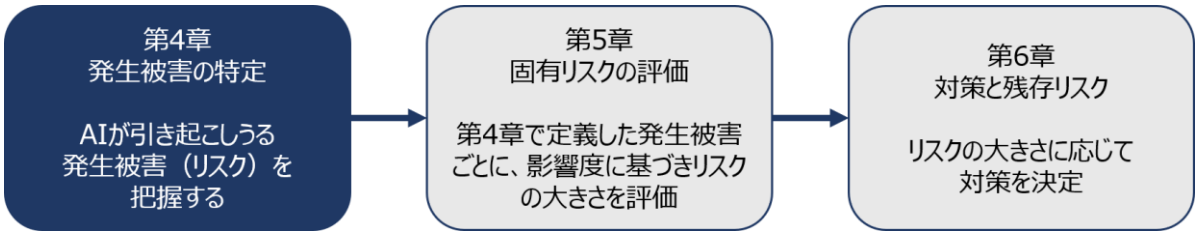


図 4-1　本章および 5,6 章との関連性

■ 発生被害の選定基準

本書における発生被害は、ユーザ企業の立場から、以下の観点に基づいて選定した。

- ユーザ企業が実際に被害を受ける立場として、業務上の実害につながる可能性が高いものを対象とした。AI 開発者・AI 提供者側に固有の問題（モデル学習工程のリスク等）は対象外とした。
- ISO/IEC 27000 が定義する情報セキュリティの 7 特性（機密性・完全性・可用性＋真正性・信頼性・責任追跡性・否認防止）に照らし、企業活動に影響を及ぼす被害カテゴリを網羅するよう選定した。
- 現時点の技術水準で顕在化しているリスクに絞り、フィジカル AI（ロボット・自動運転等）に起因する物理的事故や、将来的な人工超知能に関するリスクは対象外とした。
- 対策を講じることで発生確率を低減できる被害を優先し、技術的・運用的な対処が困難な領域（例：AI 使用による人間の知能低下）は除外した。

以上の観点から、本書では以下の 7 つの主要な発生被害を対象とする。

表 4-1　発生被害の一覧

#	発生被害	説明
A	情報の漏洩	AI システムに入力したプロンプトや AI モデル、AI システムが参照するデータベース内の情報が意図せず外部または権限のない内部ユーザに開示される。
B	付与権限外の操作	AI システムが本来与えられている権限を超えて、AI システム自身や外部の連携先システムに対して不正な実行処理を行う。
C	システムや内部データの改ざん	AI システムや RAG 等で参照する内部データベースが悪意を持って書き換えられてしまう。
D	AI システムのサービス停止	AI システムへのサイバー攻撃や依存している外部の連携先システムの停止により、AI システムが利用できなくなる。
E	虚偽や低品質な出力の利用および出力の誤用	AI システムが生成したハルシネーションや偏見、権利侵害を含む回答を利用者が妥当性を確認せずに業務やサービスに利用してしまう。
F	第三者への加害	AI システムに付与された権限の範囲内で機能が悪用されることで、自社が意図せず加害者となる。DoS[26]をはじめとする連携先システムへの直接的な攻撃や、連携先システムを踏み台にした外部のシステムへの不正アクセス、データ破壊等に悪用される。
G	過剰課金の発生	AI システムへの悪意のあるリクエストや乗っ取りにより、大幅な API 利用料やクラウド利用料が発生する。

4.2　発生被害の詳細

本節では各発生被害の概要とその発生事例を示す。

なお、本節で取り上げる事例には、実際に発生したインシデント、公開された脆弱性情報、研究者による実証・PoC、報道に基づく事例が含まれる。各事例は、発生被害の理解を目的として掲載するものであり、被害規模や発生条件の詳細は出典情報を参照されたい。また、事例名に記載した年は、原則として当該事例が公表または報道された年を示す。

[26] DoS（Denial of Service）攻撃とは、対象システムに過剰な通信や処理負荷を与えること等により、正規の利用者がシステムやサービスを利用できない状態にする攻撃のこと。

4.2.1 発生被害 A：情報の漏洩

AI システムを通じて、機密情報・個人情報・社内の非公開情報が権限のない内部ユーザや外部に流出する被害を指す。ユーザが意図せず入力プロンプトに機密情報を含めてしまった場合や、攻撃者が悪意ある指示を埋め込んだ入力プロンプトや外部コンテンツを AI モデルに読み込ませることで情報を引き出す場合に発生する。なお、RAG の参照データやモデル自体が窃取される可能性も含まれる。

事例 1	**Microsoft 365 Copilot ゼロクリック情報流出（2025 年）**[27] 【概要】 イスラエルの Aim Security 社が、Microsoft 365 Copilot にゼロクリックの脆弱性「EchoLeak」を発見した。攻撃者が細工したメールを送信するだけで、ユーザの操作なしに Copilot からチャット履歴・OneDrive 文書・Teams 会話等の機密情報が自動的に流出する状態だった。Microsoft は 2025 年 5 月に修正パッチを適用した。 【原因】 Copilot が外部から受信したメール本文に埋め込まれた悪意ある指示（間接プロンプトインジェクション）をそのまま実行してしまう設計上の問題があった。ユーザの明示的な操作を介さずに AI が外部コンテンツの指示に従ってしまうことで、情報漏洩が発生する可能性があった。

事例 2	**サムスン電子社員による ChatGPT への機密情報入力（2023 年）**[28] 【概要】 サムスン電子の社員が業務効率化のために ChatGPT を利用し、半導体の機密ソースコードや経営層の会議内容をプロンプトとして入力した。同社は事態を受け、一時的に生成 AI の利用を禁止・制限した。 【原因】 社内で生成 AI の利用ルールが整備されていない状態で、社員が個人判断で ChatGPT を業務に使用した。ChatGPT の利用規約上、入力データがモデルの学習に利用される可能性があり、入力された機密情報が自社のコントロール下から外れる結果を招いた。

[27] The Hacker News, Zero-Click AI Vulnerability Exposes Microsoft 365 Copilot Data Without User Interaction, https://thehackernews.com/2025/06/zero-click-ai-vulnerability-exposes.html

[28] Mashable, Whoops, Samsung workers accidentally leaked trade secrets via ChatGPT, https://mashable.com/article/samsung-chatgpt-leak-details

4.2.2 発生被害 B：付与権限外の操作

AI システムが本来与えられた権限の範囲を超えて、システムやデータに対して不正な操作をする被害を指す。AI エージェントが外部のシステム（メール送信・ファイル操作・API 呼び出し等）と連携する場合に顕在化しやすく、最小権限原則が守られていない環境では特にリスクが高まる。

事例 1	**AI エージェントによるメール受信トレイの全削除・機密メール外部転送（2026 年）**[29] 【概要】 OpenClaw 等の自律型 AI エージェントにメールアクセス権を与えた結果、意図せず数百通のメールが大量削除される事案や、AI の脆弱性を悪用されて機密情報が外部へ送信される事案が発生した。 【原因】 エージェントに対してメールの読み書きや削除を含む過剰な権限が付与されていた。さらに、実行前の人間の承認ステップ（Human-in-the-loop）をシステム的な制御として組み込まずプロンプトによる指示のみに依存していたため、AI の処理限界による指示忘れや、不正な指示の読み込みが直接的な被害につながった。

[29] Future US, Meta's safety director handed OpenClaw AI agents the keys to her emails — and watched it "speedrun deleting" her inbox, https://www.windowscentral.com/artificial-intelligence/meta-summer-yue-director-openclaw-ai-email-deletion

4.2.3 発生被害 C：システムや内部データの改ざん

AI システムのパラメータ（設定値）や RAG をはじめとする内部データが不正に書き換えられ、意図的に誤った出力をしてしまう被害を指す。

事例 1	**MCP サーバのツール定義改ざんによる任意コマンド実行（2026 年）[30]** 【概要】 MCP（Model Context Protocol）サーバが返すツール定義[31]を改ざんすることで、AI に誤ったパラメータや危険な操作を実行させる攻撃手法が実証された。MCP は AI エージェントが連携先システム（ファイル操作・データベース・API 等）を利用するための標準プロトコルであり、ツール定義の改ざんにより任意ファイル読み取り・任意コマンド実行・データ漏洩が可能な状態となった。 【原因】 AI エージェントが MCP サーバから受け取るツール定義を無条件に信頼する設計であったため、攻撃者がツール定義を改ざんすることで、エージェントに意図しない操作を実行させることができた。ツール定義の完全性検証が行われていなかったことが根本原因である。

[30] Charoes Huang et al., Model Context Protocol Threat Modeling and Analyzing Vulnerabilities to Prompt Injection with Tool Poisoning, https://arxiv.org/abs/2603.22489

[31] ツール定義とは、AI エージェントが外部ツール（API など）を利用できるようにするため、呼び出せる機能の名称・説明、引数の型、入力値の範囲などを構造化された仕様として記述したものである。AI エージェントはこの定義を参照し、どのツールをどのような引数で呼び出すかを判断する。

事例 2	**AI の出力を悪用した不正 SQL クエリ実行（2024 年）[32]** 【概要】 AI アプリケーション開発フレームワーク「LangChain」の特定の機能（GraphCypherQAChain）において、プロンプトインジェクションを介して深刻な SQL インジェクションが可能となる脆弱性（CVE-2024-8309）が報告された。攻撃者が悪意のある自然言語プロンプトを入力するだけで、裏側で繋がっている Neo4j などのグラフデータベースの全データ削除や、機密データの不正取得、マルチテナント環境[33]の侵害などが未認証で実行可能な状態だった。 【原因】 ユーザの自然言語をデータベース言語（Cypher[34]クエリ）に変換して直接実行する機能において、入力されたプロンプトや、LLM が生成したクエリに対する「無害化（サニタイズ）と検証」が欠落していたことが根本原因である。AI が生成したクエリを無条件で信頼し、十分な権限制限（読み取り専用権限など）を設けないままデータベースに直接投げてしまう設計だったため、プロンプトインジェクションがそのまま致命的なデータベース攻撃へと直結した。

[32] NIST, CVE-2024-8309 Detail, https://nvd.nist.gov/vuln/detail/cve-2024-8309

[33] マルチテナント環境とは、1 台の物理サーバやシステムインフラ、アプリケーションなどのリソースを、複数の企業やユーザで論理的に分割し、安全に共同利用する仕組みや環境のこと。

[34] Cypher とは、グラフデータベース用に設計された宣言型クエリ言語のこと。

4.2.4 発生被害 D：AI システムのサービス停止

AI システムが依存する外部サービス、クラウド基盤または基盤モデルの障害、DoS 攻撃や大量の不正リクエストなどにより、AI システムが正常に利用できなくなる被害を指す。AI を業務プロセスの中核に組み込んでいる企業ほど、停止時の業務影響が大きくなる。

事例 1	**Vertex AI Gemini API の障害による AI システムのサービス停止（2026 年）**[35] 【概要】 Google Cloud の Vertex AI Gemini API において、エラー率が上昇する障害が発生した。Vertex AI Gemini API は、企業が AI チャットボット、問い合わせ対応、文書要約、社内 FAQ 等に Gemini モデルを組み込む際に利用する基盤モデル API である。本障害により、Gemini API を利用する AI システムでは、回答が返らない、応答が遅い、処理が失敗するといった影響が発生する可能性があった。Google Cloud は設定変更を切り戻す等の対応を行い、サービスを復旧した。 【原因】 Gemini モデルを支える安全フィルタリングサービスへの設定変更により、一部のリクエスト処理に問題が発生した。外部基盤モデル API に対する変更管理や障害時対応の不備が、AI システムのサービス停止につながった事例である。

[35] Google Cloud Service Health Incidents, https://status.cloud.google.com/incidents/41E5S3mkTGDfkZuJZH5k

4.2.5 発生被害 E：虚偽や低品質な出力の利用および出力の誤用

AI システムが生成したハルシネーション（事実と異なるもっともらしい情報を生成すること）や偏った意見、権利侵害を含む出力を、利用者が妥当性を確認せずに業務判断やサービス提供に利用してしまう被害を指す。悪意ある攻撃に起因する場合だけでなく、LLM の確率的な生成特性そのものに起因して発生する点が他の被害と異なる。AI エージェントでは、誤った出力が次の処理ステップへ自動的に連鎖し、人間が介在しないまま影響が拡大するリスクがある。

事例 1	**ChatGPT が生成した実在しない判例等を裁判所へ提出（2023 年）**[36] 【概要】 損害賠償訴訟で、原告側弁護士が ChatGPT を用いて法律調査を行い、ChatGPT が生成した実在しない判例、虚偽の引用文、架空の判例情報を裁判書面に記載していた。裁判所は、必要な確認を行わなかったとして、弁護士 2 名と法律事務所に連帯して 5,000 ドルの制裁金などを命じた。 【原因】 ChatGPT の出力を、公式の判例データベースや裁判所記録で確認していなかった。また、相手方と裁判所から判例の実在性を疑われた後にも、原告側は ChatGPT が生成した架空の判決文も提出していた。

[36] Mata v. Avianca, Inc., No. 1:2022cv01461 - Document 54 (S.D.N.Y. 2023), https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/

4.2.6 発生被害 F：第三者への加害

AI システムがサイバー攻撃や内部不正によって掌握され、AI システムに付与された権限の範囲内で機能が悪用されることで、自社が意図せず加害者となる被害を指す。具体的には、DoS をはじめとする連携先システムへの直接的な攻撃や、連携先システムを踏み台にした外部のシステムへの不正アクセス、データ破壊等に悪用される。AI エージェントが自律的に外部 API やサービスと連携する環境では、エージェントの制御が奪われた場合に連携先への攻撃によって過負荷などが発生するリスクがある。

事例 1	**AI エージェントを悪用した大規模サイバースパイ活動（2025 年）**[37] 【概要】 2025 年 11 月、Anthropic 社は、中国系の脅威アクター（GTG-1002）が Claude Code をジェイルブレイク[38]し、サイバー攻撃エージェントとして悪用する大規模攻撃を、検知・遮断したと発表した。攻撃者は AI エージェントに「正規のセキュリティ企業の従業員として防御テストを実施している」と信じ込ませることで安全ガードレールを回避し、約 30 の組織（大手テクノロジー企業・金融機関・政府機関等）に対して偵察・脆弱性探索・認証情報の窃取・ラテラルムーブメントを自律的に実行させた。攻撃工程の 80～90%が AI エージェントによって自動実行されていた。 【原因】 AI エージェントのジェイルブレイクが成功したことで、エージェントが攻撃者の指示に従い外部組織への不正アクセスを自律的に実行した。エージェントが MCP 経由で外部ツールと連携する能力を持っていたことで、偵察からデータ窃取まで多段階の攻撃が自動化された。AI エージェントの安全ガードレールの脆弱性と、エージェントに付与された広範な外部アクセス権限が根本原因である。

[37] LAC, 大規模な AI 主導サイバー攻撃の初の報告書、アンスロピックの AI 攻撃報告が巻き起こした論争, https://www.lac.co.jp/lacwatch/report/20260202_004628.html

[38] ジェイルブレイクとは、ロールプレイ等の手法により、AI の安全機能を回避させる攻撃手法の総称である。

4.2.7 発生被害 G：過剰課金の発生

AI システムへの悪意のあるリクエストや乗っ取りまたは、意図しない処理の反復による過剰利用により、発生する金銭的被害を指す。認証情報の窃取による AI システムの不正利用でクラウド費用が爆発的に増加するケースも本被害に含まれる。

事例 1	**盗んだ認証情報によるクラウドホスト型 LLM の不正利用（2024 年）**[39] 【概要】 Sysdig の脅威リサーチチームが、盗まれたクラウド認証情報を使って AWS 等のクラウドホスト型 LLM に不正アクセスする攻撃キャンペーンを特定した。攻撃者は Claude などの AI モデルへアクセスするリバースプロキシを構築し、他のサイバー犯罪者が AI を無制限に利用できるサービスとして提供していた。被害企業は 1 日あたり最大 4 万 6,000 ドルの損失が試算された。 【原因】 企業のクラウド認証情報（API キー等）が窃取され、正規のアクセス権限を用いて LLM サービスが不正利用された。API キーの管理不備や、異常な利用量を検知する監視体制の欠如が根本原因であり、コスト上限の未設定が被害額の拡大につながった。

[39] Sysdig, LLMjacking: Stolen Cloud Credentials Used in New AI Attack, https://www.sysdig.com/blog/llmjacking-stolen-cloud-credentials-used-in-new-ai-attack

第 5 章　AI システムの固有リスク評価

AI システムを安全に活用するには、想定されるリスクを把握し、リスクの大きさに応じて、どのような対策を講じるべきかを判断する必要がある。本章では、その判断の前提となる「固有リスク評価」の考え方と評価方法を示す。

本章で評価する「固有リスク」とは、対策を講じる前の時点で、AI システムの利用目的、対象業務、扱う情報、提供範囲、連携システムの有無等から決まるリスクを指す。例えば、同じ生成 AI を利用する場合であっても、公開情報を用いた社内向けの文章作成支援と、顧客情報を参照して外部顧客へ自動応答する AI システムでは、想定される被害の大きさは大きく異なる。

情報セキュリティ分野において、リスクは一般に「発生可能性」と「影響度」の組み合わせで評価される。[40]しかし、AI システム導入の企画段階では、発生可能性を正確に見積もることは容易ではない。導入するシステムの構成、運用環境、その時々の脅威動向、そして実際に講じる対策の内容によって大きく変動するためである。

そこで本書では、AI 導入者および AI 管理者が実務において判断しやすいよう、被害が発生した場合の影響度をリスクの尺度として採用する。具体的には、第 4 章で定義した発生被害ごとに、業務的・財務的・社会的影響等の観点から固有リスクを評価する。

本章で評価した固有リスクは、第 6 章で示す対策の要否や対策水準を検討する際の判断材料となる。固有リスクを受容できない場合は複数の対策を組み合わせてリスクの低減を図る。

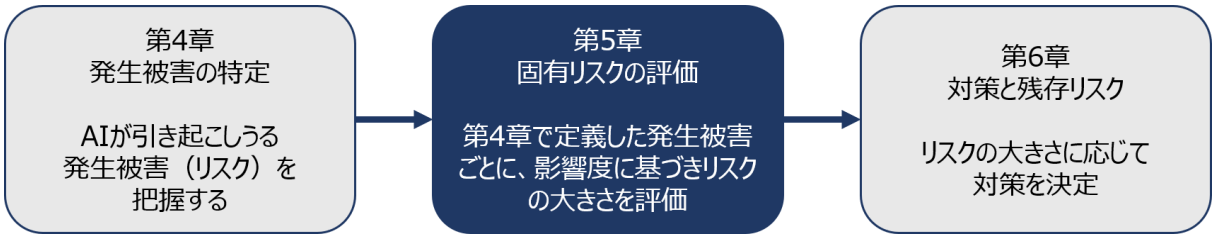


図 5-1　本章および 4,6 章との関連性

本章は次の構成で展開する。5.1 節で固有リスク評価の実施方法・体制（誰が・いつ実施するか）を、5.2 節で評価軸（何を・どう評価するか）を解説する。なお、本評価は、AI 導入時の初回評価のみならず、業務要件・提供範囲の変更時や定期点検等の運用フェーズにおける再評価でも繰り返し活用することを想定している。

[40] 独立行政法人 情報処理推進機構,リスクアセスメントの実施の手引き(NIST SP 800-30 日本語訳)
https://www.ipa.go.jp/security/reports/oversea/nist/ug65p90000019cp4-att/begoj9000000balw.pdf

5.1 固有リスク評価の実施方法・体制

本節では、固有リスク評価をいつ誰が実施するかを示す。本章の導入で述べた通り、固有リスクは、AI システムの利用目的、対象業務、扱う情報、提供範囲、外部システムとの連携状況等に応じて変化する。そのため、導入（企画）時のみではなく、AI システムのライフサイクルを通じて継続的に評価する必要がある。具体的には以下図 5-2 で示す通り、「企画」「運用前準備」「運用」の各フェーズにおいて、当該フェーズの主担当が評価を実施する。

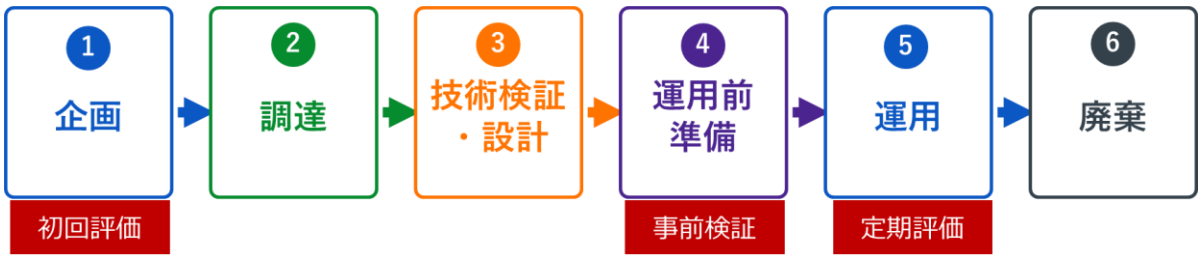


図 5-2　AI ライフサイクルにおける固有リスク評価実施タイミング

▼1.企画フェーズ（初回評価）：

AI システム活用の構想・要件定義の段階で、対象業務、扱う情報、提供範囲、連携システムの有無等を整理し、固有リスクを評価する。評価結果は、対策の検討や調達時の要求事項、設計上の制約、運用時の管理水準を検討する前提となる。

▼4.運用前準備フェーズ（事前検証）：

本番運用開始前には、技術検証・設計を経た最終的なシステム構成を踏まえて固有リスクを再評価し、企画時の評価結果と差異がないかを確認する。差異がある場合は、必要に応じて対策の見直しを行う。

▼5.運用フェーズ（定期評価）：

業務要件・利用範囲の変更時、または定期点検時に、固有リスクが変動していないかを評価する。例えば、扱うデータの機微性が高まった場合、想定外の利用が確認された場合等には再評価を行う。

■関与する部門

固有リスク評価の実施者は、AI システムのビジネス要件を把握している AI 導入者・AI 管理者を想定しているが、一部の評価にあたっては、情報システム部門等の助言を得ることが望ましい。（表 5-2 参照）また、セキュリティ部門は、各フェーズの評価結果や成果物に対するレビューや技術的助言を担う役割を想定している。

5.2 固有リスクの評価軸

本節では、第 4 章で定義した発生被害ごとに、固有リスクを評価するための評価軸を示す。評価軸は、各発生被害が発生した場合の影響度を左右する要素をもとに設定している。具体的には、AI システムの対象業務、扱う情報、提供範囲、連携システムの有無、影響を受ける主体等の観点から、業務的・財務的・社会的影響を評価する。

■評価方法

固有リスク評価は、各発生被害に対して定義された複数の評価観点について、それぞれ低・中・高の 3 段階で評価する形式で行う。低・中・高は対策を講じる前のリスクの大きさを表し、高になるほど当該被害が発生した際の影響度が大きいことを示す。各レベルの判断基準は、評価観点ごとに「レベルの判断基準例」として示す。

■固有リスクレベルの決定方法

各発生被害の固有リスクは、複数の評価観点をもとに判定する。各発生被害に対する固有リスクの大きさは、すべての評価観点を評価した上で、最も高いレベルをその発生被害に対する固有リスクの大きさとする。これは、いずれか 1 つの観点でも「高」と評価される要素があれば、当該被害が発生した際の影響度は「高」となりうるためである。

なお、いずれかの評価観点で「高」と判定された場合でも、途中で評価を打ち切らず、すべての観点を評価する。各発生被害の総合評価が同じ「高」であっても、どの観点が高いか、また高と評価された観点がいくつあるかによってリスクの性質や影響の広がりが異なるため、この評価結果が第 6 章における対策の選択の判断材料となる。

■評価時の留意点

各評価観点に示すレベルの判断基準例は一般的な目安であり、自社の業界特性、業務内容、保有情報の性質等に応じて、調整して活用することが望ましい。例えば「扱う情報の機密性」における「高」の例として個人情報・機密情報を挙げているが、医療・金融等の業界においては、より細かい区分や独自の機密度区分を適用することも想定される。なお、対策予定の有無によって、固有リスクを低く見積もらないよう留意する。

■評価観点における評価対象範囲の考え方

評価観点によっては、AI システム単体だけでなく、AI システムと連携するシステム（連携先システム）や、連携先を介して影響を受けるシステム（間接影響システム）、同一ネットワーク上のシステム、社外の連携先システムまで含めて評価する必要がある。図 5-3 は、評価対象に含めるシステム範囲の考え方を示したものである。各発生被害の評価にあたっては、必要に応じて本図および定義表を参照し、評価対象となる範囲を確認する。

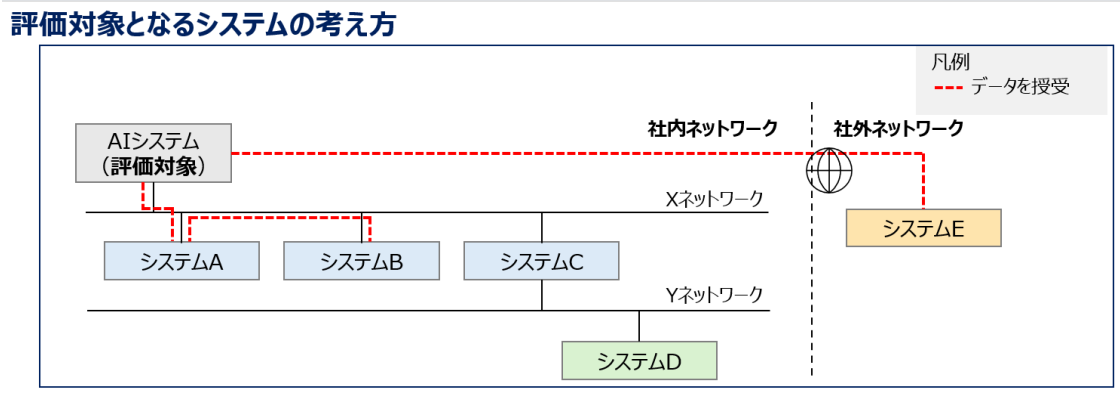


図 5-3　評価対象となるシステム範囲の考え方

表 5-1　評価対象となるシステム範囲定義

名称	定義	図 5-3 における対象システム
連携先システム	AI システムが直接データ等を授受しているシステム。 （例：AI システムが人事システム（システム A）とデータ連携をしている。）	A
間接影響システム	AI システムと直接連携していないが、連携先システムを介して間接的に影響を受けうるシステム。 （例：AI システムが人事システム（システム A）に誤ったデータを出力した場合、人事システムと連携する給与システム（システム B）にも誤りが波及する。）	B
同一ネットワーク上のシステム	AI システムと同じネットワークに存在するすべてのシステム。	A,B,C
社外連携先システム	AI システムがデータを授受している社外のシステム。 （例：AI システムが取引先システム（システム E）とデータ連携している。）	E

■発生被害ごとの評価観点（全体一覧）

各発生被害に対する評価観点の全体像を以下に示す。詳細は 5.2.1 節から 5.2.7 節にて発生被害ごとに記述する。なお、実施者は 5.1 節の関与する部門の通りだが、特にシステム構成、ネットワーク環境、権限設定、連携先システムなど社内の情報システムに関する知識が必要となる評価観点については、表 5-2 で[情報システム部門の支援要否]として示す。

表 5-2　発生被害ごとの評価観点一覧

発生被害	評価観点	情報システム部門の支援要否
A. 情報の漏洩	1.扱う情報の機密性 / 2.AI システムの提供範囲	
B. 付与権限外の操作	1. 影響が及ぶ可能性のあるシステムの重要度 / 2. 実行される可能性のある操作の重大性 / 3.影響を受ける主体	○
C. システムや内部データの改ざん	1. 改ざんによる業務への影響度 / 2.影響を受ける主体 / 3. 改ざん情報の伝播範囲	○
D. AI システムのサービス停止	1.停止による影響度/ 2.影響を受ける主体 / 3.事業影響を考慮した許容停止時間	
E. 虚偽や低品質な出力の利用および出力の誤用	1. 誤りによる業務への影響度 / 2.影響を受ける主体	
F. 第三者への加害	1.加害の及ぶ範囲 / 2.影響が及ぶ可能性のあるシステムの重要度	○
G. 過剰課金の発生	1.AI システムの処理回数/ 2.処理対象のデータサイズ / 3.AI システムの提供範囲	

【凡例】

(情報システム部門の支援要否)：○は情報システム部門の支援が望ましい、空欄は AI 導入者・AI 管理者のみでも評価可能であることを示す。

5.2.1 発生被害 A：情報の漏洩

表 5-3 発生被害 A「情報の漏洩」の評価軸（発生被害の定義は 4.2.1 節参照）

評価観点	観点の説明	レベル	レベルの判断基準例
1. 扱う情報の機密性	AI システムが扱う情報（入力・処理・保存・参照する情報）の機密性を評価する。機密性が高い情報ほど、漏洩した際の被害が大きくなる。	低	公開情報（例：Web サイト掲載情報、公開仕様書）
		中	社内情報（例：社内マニュアル、社内規程）
		高	個人情報（例：氏名、住所、連絡先）、機密情報（例：未公開の経営情報、契約情報）
2. AI システムの提供範囲	AI システムを利用できる対象者の範囲を評価する。提供範囲が広いほど、情報漏洩時に拡散範囲が広がるため、被害が大きくなる。	低	社内のみ
		中	社外の限定的な相手（例：特定の取引先、委託先）
		高	不特定多数（例：インターネット上に AI システムを公開）

5.2.2 発生被害 B：付与権限外の操作

■評価の前提

本発生被害は、AI システムが連携するシステムや構成に応じてリスクが変わる被害である。評価にあたっては、「AI には閲覧権限しか与えない」といった対策実施後の状態ではなく、AI システムが不正に操作された場合に、どのシステムへ影響がおよび得るか、どのような操作が実行され得るか、誰に影響がおよび得るかを評価する。

■評価範囲の留意点

本発生被害では、連携先システムだけでなく、間接影響システムや、同一ネットワーク上のシステムも考慮する。なお、同一ネットワーク上に存在していても、通信経路が制御され、AI システムから到達できないシステムは評価対象外としてよい。

「B.付与権限外の操作」における評価範囲

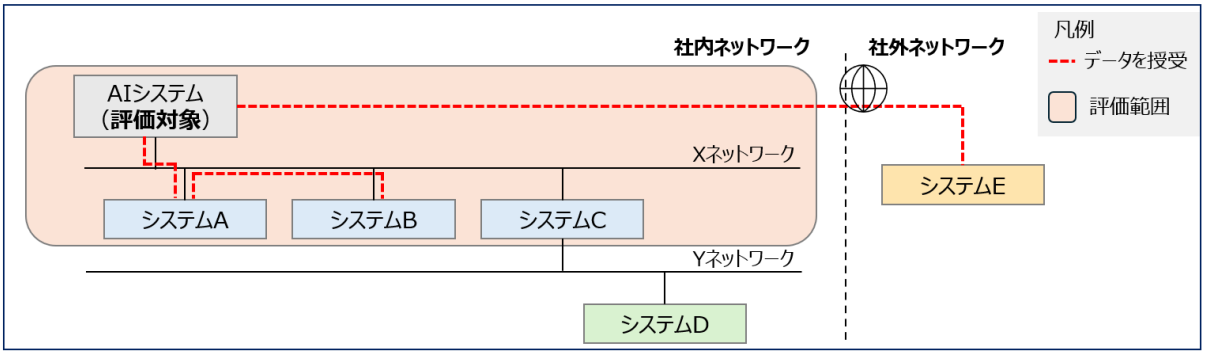


図 5-4　発生被害 B　付与権限外の操作における評価範囲

表 5-4 発生被害 B「付与権限外の操作」の評価軸（発生被害の定義は 4.2.2 節参照）

評価観点	観点の説明	レベル	レベルの判断基準例
1. 影響が及ぶ可能性のあるシステムの重要度	AI システムが不正に操作された場合に、影響が及ぶ可能性のあるシステム（同一ネットワーク上のシステム）の重要度を評価する。影響が及ぶ可能性のあるシステムが複数ある場合は、最も重要度の高いシステムのレベルで評価する。	低	業務やサービスへの軽微な影響（例：社内 Wiki、非重要データ）
		中	業務やサービスの停止・大幅な遅延（例：案件管理、在庫管理、社内業務システム）
		高	事業継続の危機や深刻な法令違反に直結または同一ネットワーク上のシステムへの侵害により社外システムに影響を与える（例：顧客情報管理、契約管理、基幹システム）
2. 実行される可能性のある操作の重大性	影響が及ぶ可能性のあるシステム（同一ネットワーク上のシステム）において、AI システムが不正操作された場合に実行される可能性のある操作の重大性を評価する。実行される可能性のある操作が複数ある場合は、最も重大な操作を基準に評価する。	低	閲覧・参照レベルの操作
		中	復元可能な変更・削除操作（攻撃者により更新・登録・削除はされるが、バックアップ等から復元できる）
		高	復元不可能な変更・削除操作、または同一ネットワーク上のシステムへの侵害により社外へも影響が及ぶ操作（例：データの物理削除、接続先変更・追加、本番システムの設定変更、外部への発信・決済）
3. 影響を受ける主体	AI システムが不正に操作された場合に、影響が及ぶ可能性のあるシステム（同一ネットワーク上のシステム）において、最も影響を受ける可能性がある主体を評価する。	低	特定のチーム・部門、個人
		中	自社全体、複数の部門
		高	一般顧客、取引先、社会全体

5.2.3 発生被害 C：システムや内部データの改ざん

■評価の前提

システムや内部データの改ざんでは、改ざんされた対象そのものの重要度に加え、その情報が誰に影響し、後続業務や連携先システム、間接影響システムへの波及度合いを評価する。

表 5-5 発生被害 C「システムや内部データの改ざん」の評価軸
（発生被害の定義は 4.2.3 節参照）

評価観点	観点の説明	レベル	レベルの判断基準例
1. 改ざんによる業務への影響度	RAG や設定値等の AI システムが参照するデータが改ざんされ、誤った出力によって業務に与える影響を評価する。	低	誤りがあっても業務への影響が小さい（例：業務上の参考情報）
		中	誤りが業務に影響するが、特定の範囲に限定される（例：業務データ、案件情報）
		高	誤りが意思決定・顧客・財務に重大な影響を及ぼす（例：基幹データ、顧客情報、機密情報）
2. 影響を受ける主体	RAG や設定値等の AI システムが参照するデータが改ざんされ、誤った出力によって影響を受ける主体を評価する。	低	特定のチーム・部門、個人
		中	自社全体、複数の部門（例：全社共通ポータル、全社向け AI ボット）
		高	一般顧客、取引先、社会全体（例：顧客向け Web サイト、顧客情報）
3. 改ざん情報の伝播範囲	改ざんされた情報が、連携先システムや、間接影響システム、後続の業務へ伝播する構成やプロセスになっているかを評価する。 ※参考：図 5-5	低	連携先システム無し、かつ AI システムが業務プロセスに組み込まれない（例：個人の思考整理・アイデア出し、一時的な調査）
		中	連携先システムの有無を問わず、AI システムの出力が人の手を介して後続業務に使われる（例：AI システムが出力した数値を人が別システムに入力する、AI システムの出力を社内向け報告書に転記する）
		高	連携先システムに改ざんされた情報が送信され、間接影響システムや社外システムにも影響が及ぶ仕組み。

「C.システムや内部データの改ざん」の評価イメージ（3.改ざん情報の伝播範囲：中・高レベル）

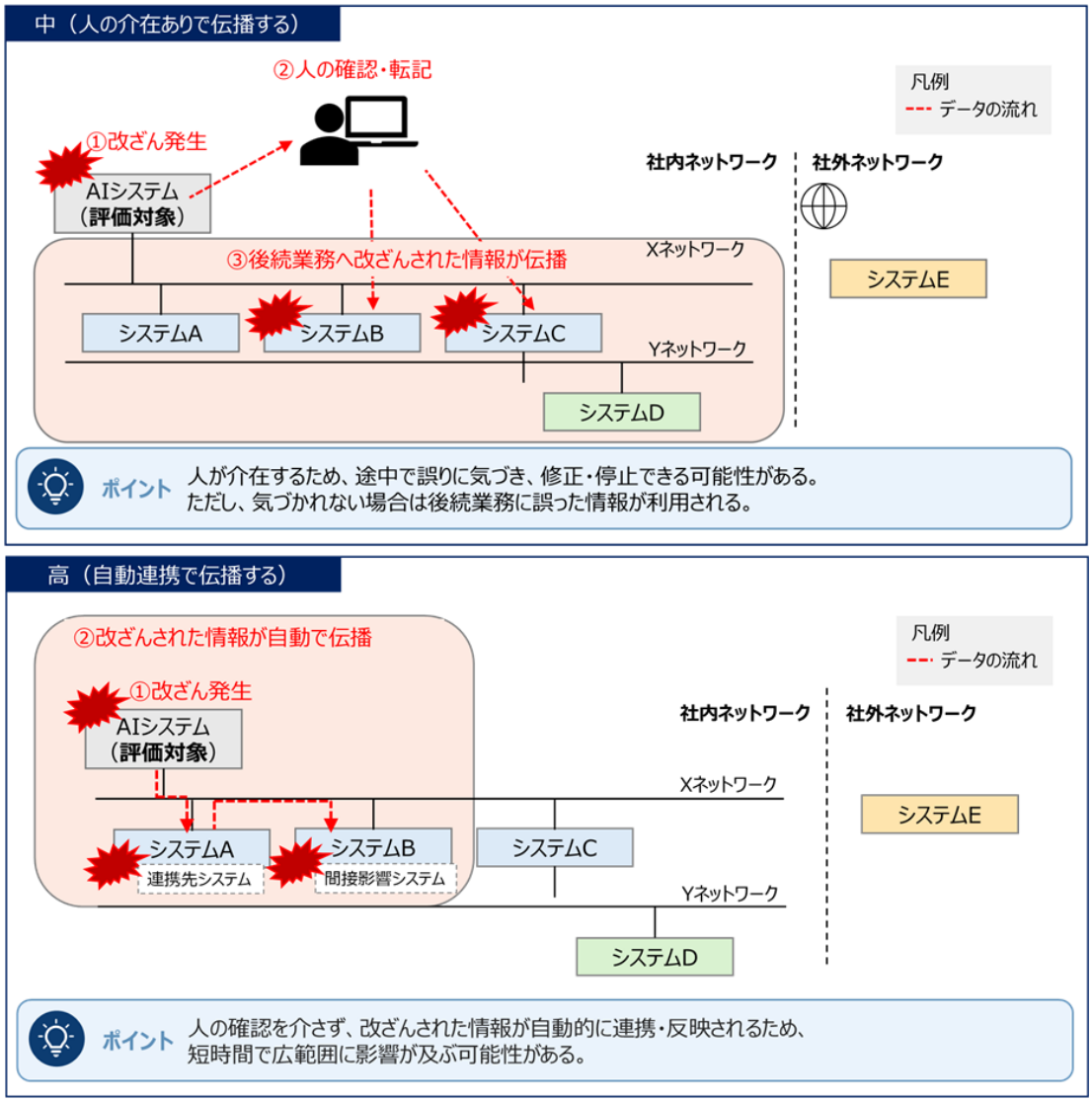


図 5-5　改ざん情報の伝播範囲の考え方

5.2.4 発生被害 D：AI システムのサービス停止

表 5-6 発生被害 D「AI システムのサービス停止」の評価軸（発生被害の定義は 4.2.4 節参照）

評価観点	観点の説明	レベル	レベルの判断基準例
1. 停止による影響度	AI システムが停止した場合の業務やサービスへの影響度を評価する。	低	AI システムが業務プロセスに組み込まれておらず、使用できなくても業務は進行可能（例：個人のアイデア出しや壁打ち、簡単な調査等）
		中	AI システムが業務プロセスに組み込まれ、停止時に業務やサービスに遅延は生じるが、手作業等で代替可能（例：社内向け文書の自動要約、問い合わせの一次対応）
		高	AI システムが業務プロセスに組み込まれ、手作業での代替が不可能。または、AI システムが業務・サービスの根幹を担う。（例：AI による大量データの解析、金融取引判定）
2. 影響を受ける主体	AI システムが停止した際に、影響を受ける主体を評価する。	低	特定のチーム・部門、個人
		中	自社全体、複数の部門（例：全社共通ポータル、全社向け AI ボット）
		高	一般顧客、取引先、社会全体（例：顧客向け Web サイト、公共サービス向け AI）
3. 事業影響を考慮した許容停止時間	AI システムが停止した際に、損失、追加対応コスト等の影響を踏まえて、許容できるサービス停止時間を評価する。	低	事業影響は軽微であるため、複数日以上の停止が許容可能
		中	業務遅延、追加対応コスト、限定的な影響が発生するため、数時間～1 日程度での復旧が求められる。
		高	重大な売上損失、広範囲な顧客影響、社会的信用低下が発生し得るため、数時間以内の復旧が求められる。

5.2.5 発生被害 E：虚偽や低品質な出力の利用および出力の誤用

表 5-7 発生被害 E「虚偽や低品質な出力の利用および出力の誤用」の評価軸
（発生被害の定義は 4.2.5 節参照）

評価観点	観点の説明	レベル	レベルの判断基準例
1.誤りによる業務への影響度	AI システムの出力をもとに意思決定や業務を行う際、誤りや品質不備が業務・法令・信用に与える影響の大きさを評価する。	低	AI システムが業務プロセスに組み込まれないため、業務への影響がほぼない。誤りがあっても修正が容易で、外部への影響もない。（例：社内メモ、メール下書き、アイデア出し）
		中	業務品質や効率に影響が出る。修正コストや手戻りが発生するが、法令・契約違反には至らない。（例：社内報告書、顧客提案書のたたき台）
		高	法令違反、契約違反、または重大な信用低下につながる可能性があり、重大な損失を招く。（例：法務判断、契約内容の確定、融資判断、顧客への正式回答）
2. 影響を受ける主体	虚偽や低品質な出力を最終的に受け取る、あるいはその影響を受ける主体を評価する。	低	特定のチーム・部門、個人
		中	自社全体、複数の部門（例：全社共通ポータル、全社向け AI ボット）
		高	一般顧客、取引先、社会全体（例：顧客向け Web サイト）

5.2.6 発生被害 F：第三者への加害

表 5-8 発生被害 F「第三者への加害」の評価軸（発生被害の定義は 4.2.6 節参照）

評価観点	観点の説明	レベル	レベルの判断基準例
1. 加害の及ぶ範囲	AI システムが乗っ取られた際に、加害が及ぶ可能性がある範囲を評価する。 ※参考：図 5-6	低	連携先システムが存在しない、独立した環境。
		中	評価対象 AI システムが、社内のシステムに接続する環境。（同一ネットワーク上のシステムが全て社内で閉じている。）
		高	評価対象 AI システムが、社外のシステムに接続する環境。
2. 影響が及ぶ可能性のあるシステムの重要度	自身の AI システムが乗っ取られた際に、影響が及ぶ可能性のあるシステム（例：同一ネットワーク上のシステムや連携先社外システム）の重要度を評価する。影響が及ぶ可能性のあるシステムが複数ある場合は、最も重要度の高いシステムのレベルで評価する	低	業務やサービスへの軽微な影響（例：社内 Wiki、非重要データ）
		中	業務やサービスの停止・大幅な遅延（例：案件管理、在庫管理、社内業務システム）
		高	事業継続の危機や深刻な法令違反に直結（例：顧客情報管理、契約管理、基幹システム）

「F.第三者への加害」の評価イメージ（1.加害の及ぶ範囲：中・高レベル）

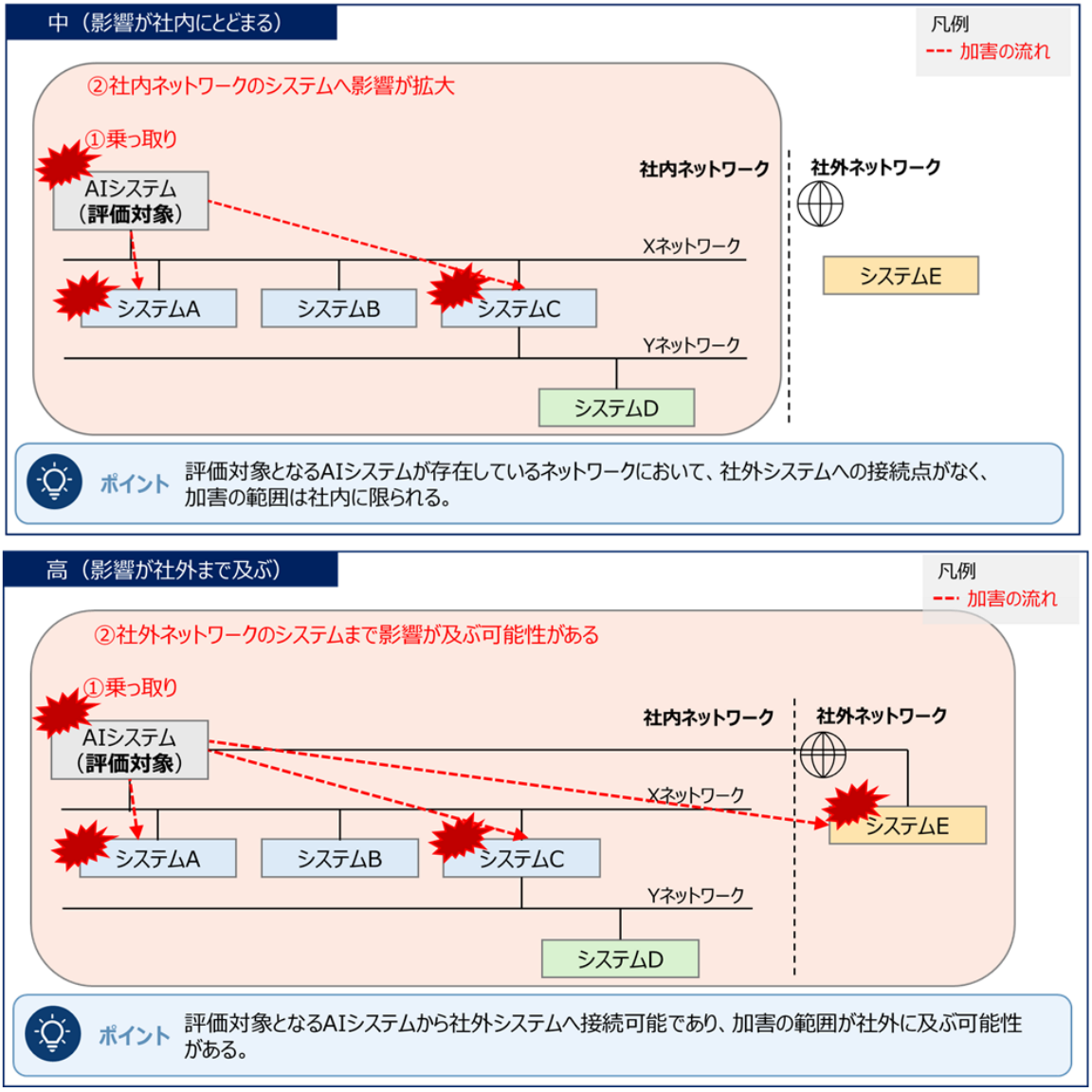


図 5-6　加害の及ぶ範囲の考え方

5.2.7 発生被害 G：過剰課金の発生

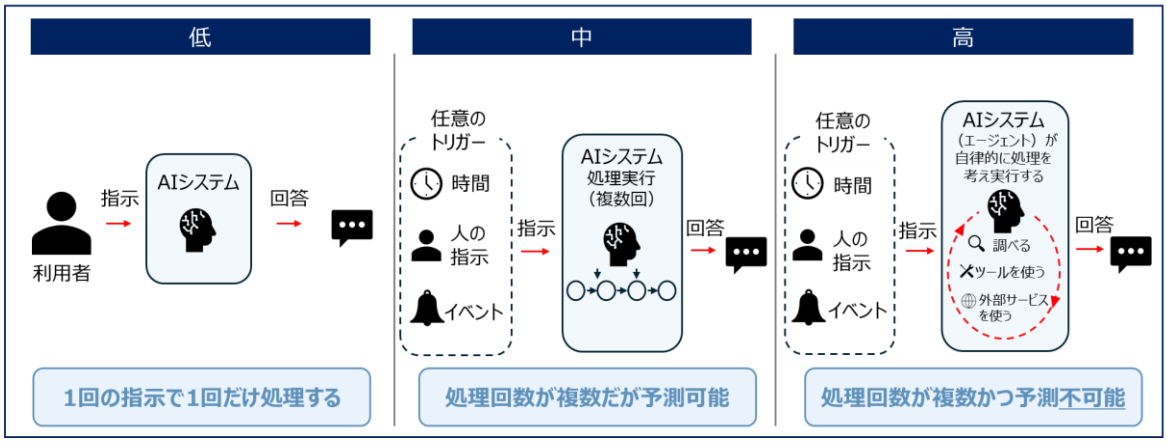
表 5-9 発生被害 G「過剰課金の発生」の評価軸（発生被害の定義は 4.2.7 節参照）

評価観点	観点の説明	レベル	レベルの判断基準例
1. AI システムの処理回数	AI システムの動作条件と 1 イベントあたり何回 AI モデルが処理を行う仕組みかを評価する。処理回数が多い、または AI システムが自律的に処理を繰り返す仕組みであるほど、過剰課金が発生しやすくなる。 ※参考 図 5-7	低	1 回の人間の指示で、処理回数が 1 回である。（例：翻訳、メール要約、単発の文章作成）
		中	時間やイベント（人間の指示を含む）といった任意のトリガーで AI システムが処理を開始し、その処理回数が予測可能である。（例：API で AI モデルを呼び出すシステム、問合せチャットボット）
		高	時間やイベント（人間の指示を含む）といった任意のトリガーで AI システムが処理を開始し、その処理回数が予測不可能である。（例：オーケストレータ[41]を用いた AI エージェント）
2.処理対象のデータサイズ	AI システムが探索・処理するデータサイズを評価する。処理対象のデータサイズが大きい、または予測が見込めない場合、過剰課金が発生しやすくなる。 ※参考 図 5-7	低	入力されたデータのみを処理し、RAG や連携先システムにデータ探索を行わない。（例：メール作成、コード修正）
		中	入力されたデータの他、RAG や連携先システムのデータベース・ファイル群にデータ探索を行う。（例：社内文書検索）
		高	入力されたデータの他、Web 上の情報やリアルタイムデータなどを対象に探索を行う。処理対象が広範または流動的で、処理量を事前に見込みにくい。 （例：Web 上のトレンド監視、リアルタイム市場データ分析）

[41] オーケストレータとは、複数の AI エージェントや連携先システム、外部 API を統合し、タスクの分解・割り当て・実行順序の制御・結果の統合など、全体のワークフローを管理する機能のこと。

評価観点	観点の説明	レベル	レベルの判断基準例
3. AI システムの提供範囲	AI システムに指示を出せる人の範囲を評価する。利用範囲が広いほど、大量利用や想定外の利用によって過剰課金が発生しやすくなる。	低	特定のチーム・部門、個人（利用人数が限定的、利用者の特定管理が容易である）
		中	社内全体・複数の部門
		高	外部、不特定多数（例：インターネット上に AI システムを公開）

「G.過剰課金の発生」の評価イメージ（1. AIシステムの処理回数）



「G.過剰課金の発生」の評価イメージ（2.処理対象のデータサイズ）

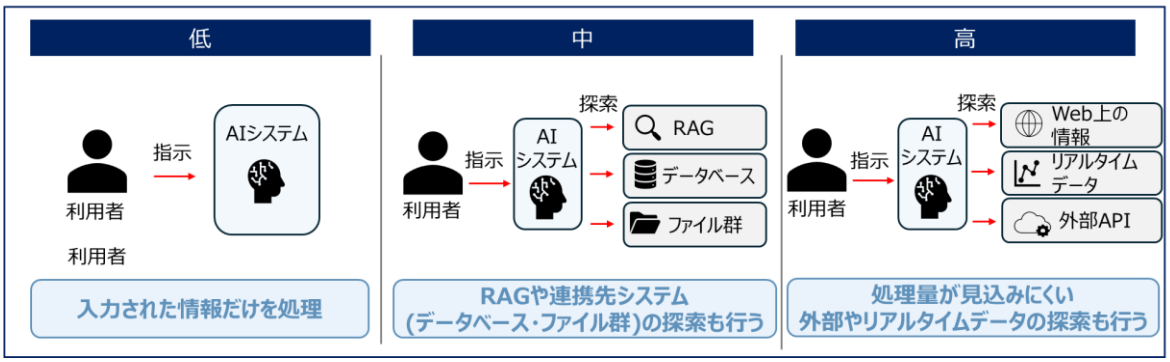


図 5-7　AI システムの処理回数 および 処理対象のデータサイズの考え方

第 6 章　対策の効果と残存リスク

本章では、第 4 章で定義した発生被害の発生確率を低減するためのセキュリティ対策を体系的に整理し（6.1 節）、各発生被害と対策の関係を明らかにする（6.2 節）。各発生被害に対する対策の具体的な組み合わせ方、優先順位、残存リスク、導入・運用上の留意点については、付録「発生被害ごとの対策の進め方」に詳述する。

6.1 節では、AI 導入者や AI 管理者が実施すべき対策を 3 つのカテゴリに分類して提示する。まず「技術要件（T）」として、AI 提供者に対して調達・契約段階で求めるべきセキュリティ要件を示す。次に「運用対策（O）」として、AI 管理者が自社の環境で設定・実施すべき対策を示す。最後に「人的統制（H）」として人間の判断・承認により技術や運用ではカバーしきれないリスクを補完する対策を示す。

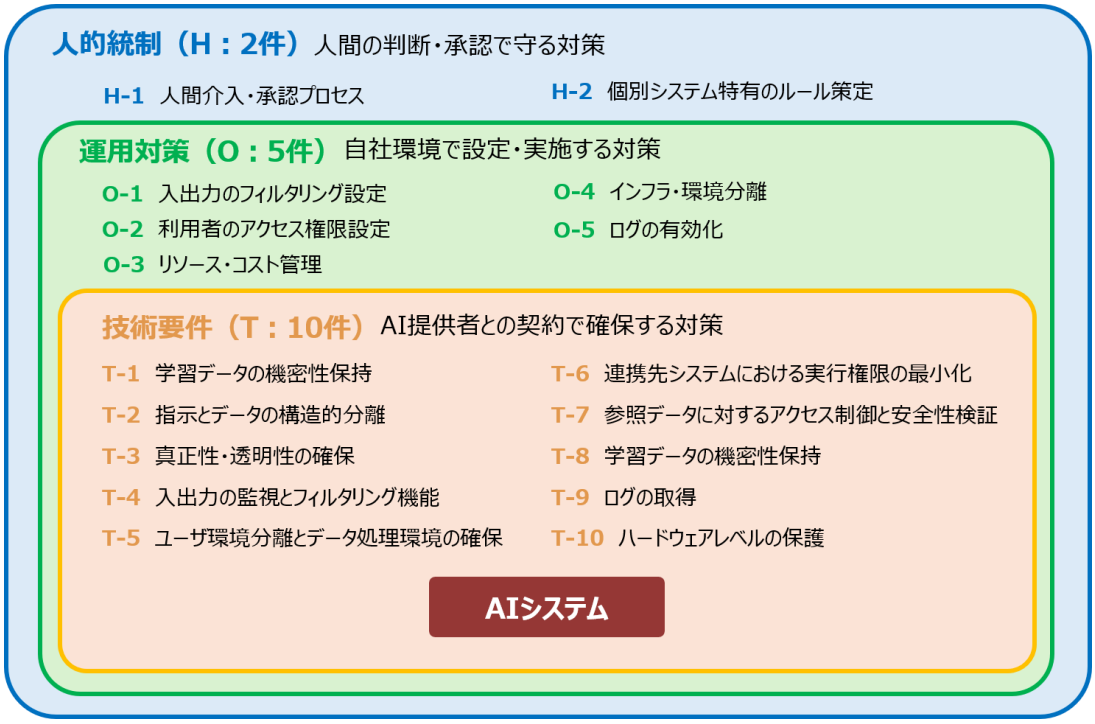


図 6-1 3 つの対策カテゴリによる多層防御の構造

6.2 節では、第 4 章で定義した 7 つの発生被害（A～G）と 6.1 節の対策との対応関係をマトリクス表として示す。第 5 章で評価した固有リスクの高さやリソース制約に基づき、自社に必要な対策を把握するための全体像として活用されたい。

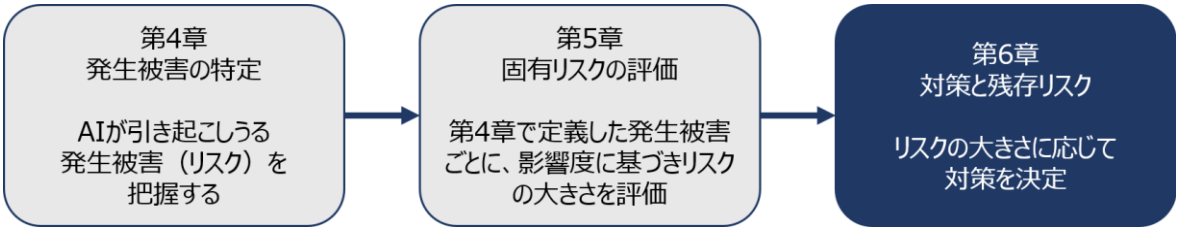


図 6-2　本章および 4,5 章との関連性

なお、第 5 章で評価した固有リスクの高さや評価基準、評価したリスクの許容度は、講じる対策を選定する目安となる。固有リスクが許容水準を大きく上回る場合、または被害の影響が深刻な場合は、技術要件を中心に運用対策・人的統制を組み合わせ、単一の対策に依存しない多層的な防御を構築することが望ましい。そうでない場合で、リスク評価が中程度の場合は、自社のリソース制約を踏まえた上で優先度の高い対策を選択し、講じられない対策については残存リスクとして明示・管理する。低の場合は、最低限の人的統制を講じた上で残存リスクとして受容することも合理的な選択となる。

また、リスクへの対応方法はセキュリティ対策によるリスク低減に限らない。業務プロセスの見直しや該当用途での AI 利用範囲を制限することによるリスク回避や、AI 保険の活用によるリスク移転（2.4.5 節）も選択肢となりうる。本章では主にリスク低減の手段としての対策を扱うが、固有リスクの高さや自社の状況によっては、これらの対応方法を組み合わせることも検討すべきである。

6.1 対策の一覧

本節では、AI システムを安全に導入・運用するために必要なセキュリティ対策を、実施主体と対策の性質に基づき「技術要件」「運用対策」「人的統制」の 3 つに分類して提示する。各分類の対策は独立したものではなく、相互に補完し合うことで初めて実効的な防御が実現されるため、6.2 節（付録を含む）と自社の状況を考慮して活用されたい。

6.1.1 技術要件（T）

本項では、AI 導入者が企画・調達する際、および AI 管理者がシステムを更新する際に、AI 提供者に対して提示すべきセキュリティに関する技術要件を示す。

AI のセキュリティ対策には、モデルの内部構造や学習プロセスに依存するものが多く、ユーザ企業だけでは実装できない対策が存在する。例えば、学習データからの情報漏洩を防ぐ仕組みや、プロンプトインジェクション攻撃への耐性は、AI 提供者が組み込むべき技術である。また、AI 開発者においても、モデルの安全性に関する対策だけでなく、AI のガバナンスに関する機能（AI エージェントの ID 管理、ログの取得、権限管理など）も整備が進みはじめている。これらの要件を AI システム導入後に追加することは、システム構成変更を伴うため、現実的ではない。そのため、導入前の企画・調達フェーズで RFP（提案依頼書）や調達仕様書に明記し、契約条件として合意しておくことが重要である。

以下に AI 提供者に求めるべき 10 の技術要件項目を示す。ただし、これらの項目には技術的難易度が高いものや研究段階のものが含まれており、全てを実装・要求できるとは限らない。また、クラウド型の SaaS として AI を導入する場合、提供者側の仕様に依存するため、個別の要件を契約条件として交渉できないケースもあることに留意する。その場合、AI 提供者が第三者機関による AI 脆弱性診断を受けているかを確認し、その結果をもって技術要件への適合性を間接的に評価することも実務上有効な手段である。

（1） **学習データの機密性保持**（T-1）

LLM は、学習データに含まれる情報を推論時に再現してしまう場合がある。AI 提供者のモデルが自社または他社の機密情報を学習データとして取り込んでいた場合、推論結果を通じてそれらの情報が漏洩

するリスクがある。

AI 提供者に対して、採用する AI モデルが、データに特殊なノイズを混ぜるなどして特定の学習データを特定・復元できなくする安全設計（差分プライバシー[42]）が適用されていることの確認や開示を求めることで、学習データ経由の情報漏洩リスクを低減する。ただし、差分プライバシーは AI モデルの設計・学習工程に関わるものであり、AI 開発者側の対応が必要となるため、ユーザ企業が AI 提供者に対して直接要求・確認することが困難な場合がある。その場合の現実的な対応として Zero Data Retention（入力データを AI 提供者側で一切保持・学習利用しないことを保証する契約条件）を契約に盛り込むことが有効である。これにより、自社が入力したデータが将来のモデル学習に取り込まれ、他のユーザへの回答を通じて情報が漏洩するリスクを契約上排除できる。

（2） **指示とデータの構造的分離**（T-2）

プロンプトインジェクション攻撃は、外部データに紛れ込ませた不正な指示を AI に実行させるものであり、情報漏洩や権限外操作などの多くの発生被害の起点となる。この攻撃は、AI がシステムの指示とユーザ入力・外部データを内部的に区別できないことに起因する。

システム指示と入力データを構造的に分離するアーキテクチャ（開発者の命令を最優先する設計や、指示とデータを分けて認識させる専用のデータ形式等）は、プロンプトインジェクション攻撃への対策として研究が進んでいる。現時点では、多くの AI モデルがこの分離を十分に実現できておらず[43]、AI システムに広く組み込まれるに至っていない。引き続き研究動向を注視し、AI 提供者が当該機能を提供する場合には積極的に活用することが望ましい。

（3） **真正性・透明性の確保**（T-3）

AI が生成したコンテンツがディープフェイクや偽情報として悪用されるリスクや、利用しているモデルのサプライチェーンに脆弱性が潜むリスクがある。インシデント発生時に「どのモデルが、どのデータを用いて、何を生成したか」を追跡できなければ、原因究明と再発防止が困難になる。

AI 提供者に対して、以下を求めることで、生成物の真正性確認とモデルの安全性監査を可能にする。

---

[42] 差分プライバシーは、学習データに含まれる個人・機密情報がモデル出力から復元されるリスクを数学的に低減する手法である。AI モデルへの適用が徐々に進んでおり、一部の AI 開発者では実用化されている。一方で、プライバシー保護強度を高めるほどモデル性能が低下するトレードオフが存在する。参考：Michele Miranda et al., "Preserving Privacy in Large Language Models: A Survey on Current Threats and Solutions", 2025, https://arxiv.org/abs/2408.05212

[43] Egor Zverev et al., "Can LLMs Separate Instructions From Data? And What Do We Even Mean By That?", ICLR 2025, https://arxiv.org/abs/2403.06833, テストした全 LLM が指示・データの信頼できる分離に失敗し、プロンプトエンジニアリングやファインチューニング等の緩和策でも不十分であることが示されている。

- AI による生成物であることを証明するメタデータの埋め込み機能（電子透かし[44]や C2PA[45]等の来歴証明技術）
- AI システムの開発・運用に関わる構成要素の一覧表（SBOM）の提出
- AI モデルの構成情報（使用ライブラリ、学習データの出所等）の開示

AI モデルの構成情報の開示は、商用利用可能な AI モデルでは困難であるが、オープンソースの AI モデルを活用して、チューニングをする場合は把握できる可能性がある。

（4） **入出力の監視とフィルタリング機能**（T-4）

AI モデル自体の安全対策だけでは、すべての攻撃や不適切な出力を防ぐことは困難である。モデルの回答が一見正しく見えても、機密情報の断片が含まれていたり、不適切な内容が紛れ込んでいたりする場合がある。また、AI モデルをクラウド API 経由で利用する場合、プロンプトに含まれるデータは AI 提供者のサーバに送信されるため、自社の個人情報や機密情報が意図せず外部に公開されるリスクもある。

AI 提供者に対して、以下の機能の実装を求めることで、モデルの誤動作や攻撃がすり抜けた場合の多層的な防御を補強することができる。ただし、これらも万全ではなく、単独での完全な防御を保証するものではない。[46]

- モデルとは独立した入出力のリアルタイム検閲・遮断機能（ガードレール）
- プロンプト内の個人情報・機密項目を自動で検知・匿名化する動的マスキング機能[47]

[44] 電子透かしとは、コンテンツに識別情報を知覚されにくい形で埋め込む技術のこと。ファイル形式の変換や、トリミング等の加工を経ても消えにくい特性を持つ。

[45] C2PA とは、コンテンツの作成者・作成日時・使用ツール等の来歴情報を署名付きメタデータとしてファイルに付加する国際的なオープン標準仕様のこと。改ざんの検知にも利用できる。

[46] OWASP Top 10 for LLM Applications 2025, LLM01:2025 Prompt Injection では「生成 AI の確率的な性質に起因するため、プロンプトインジェクションの確実な防御策が存在するかは不明（it is unclear if there are fool-proof methods of prevention for prompt injection）」と明示しており、完全防止は現時点で保証できないことを示す。https://genai.owasp.org/llmrisk/llm01-prompt-injection/

[47] 動的マスキングには固有表現認識（Named Entity Recognition: NER）等の技術が用いられる。NER は曖昧語・多義語・表記ゆれ等に対する検知精度に限界があり、すべての個人情報や機密情報を漏れなく検出できるとは限らない。Devansh Singh, Sundaraparipurnan Narayanan, "Unmasking the Reality of PII Masking Models: Performance Gaps and the Call for Accountability", 2025, https://arxiv.org/abs/2504.12308

- データラベル（機密度・人間によるレビュー状況・AI 入力可否等を示すラベル）を読み取り、ラベルに基づいてフィルタリングルールを自動適用する機能

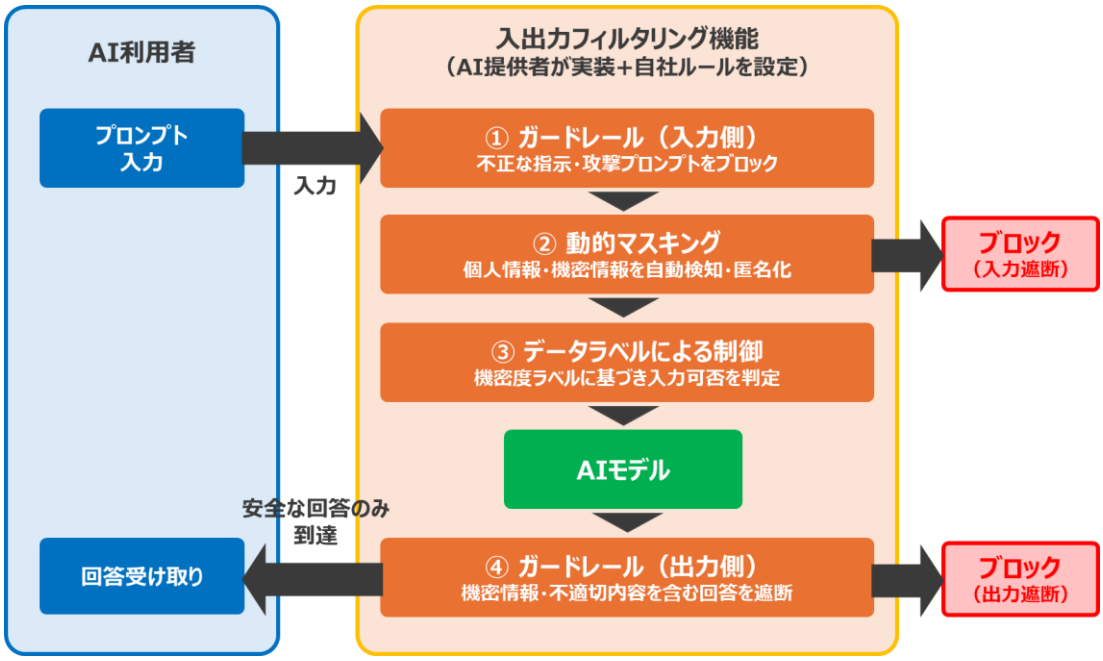


図 6-3　入出力の監視とフィルタリングの構成

（5）　**ユーザ環境分離とデータ処理環境の確保**（T-5）

複数の利用者が同一の AI 基盤を共有する環境（マルチテナント環境）では、あるユーザの入力データや会話履歴が別のユーザの応答に混入するリスクがある。また、AI システムをクラウドサービスとして利用する場合、自社のデータがどの地域のサーバで処理・保存されるかによって、越境データ移転に関する法令上の制約を受ける場合がある。

以下の点について導入前に AI 提供者と合意しておくことが重要である。

- ユーザやセッションごとの実行環境の論理的分離
- AI 提供者の提供形態（クラウド/オンプレミス/ハイブリッド）の選択肢
- データが処理・保存される地理的な場所（国内／海外）および越境データ移転に関する法令上の要件への適合

扱うデータの機密性が高い場合は、ローカル LLM（オンプレミスやクローズドな自社環境でのモデル運用）も選択肢となる。適切に設計・運用すれば、データの外部送信を抑制できるため、情報漏洩リスクを大幅に低減できる一方、ローカル LLM 自体の防御負荷が発生するため、扱うデータの機密性とコストのバランスを踏まえて判断する。

（6）　**連携先システムにおける実行権限の最小化**（T-6）

AI エージェントは連携先システムと連携して動作するが、過大な権限が付与されていると、攻撃者に AI システムが乗っ取られた場合に連携先システムの破壊やデータの不正操作といった重大な被害に直結する恐れがある。こうしたリスクには、Zero Trust 原則（何も信頼せず常に検証するというセキュリティ設

計思想）を適用することが求められる。[48]

AI 提供者に対して、AI エージェントのオーケストレータやエージェント自身の権限を最小限に抑え、不正なコマンド実行が構造的に不可能な設計を求めることで、AI システムが侵害された場合の被害範囲を限定する。具体的には、データベースをはじめとする連携先システムとの接続を閲覧（読取専用）権限のみに制限する設計や、データベースを不正操作するサイバー攻撃（SQL インジェクション）を防ぐ安全なデータ処理方式（パラメータ化クエリ）および AI システムが利用できる外部ツールの機能や入力範囲を厳しくする設計（ツール定義の厳密化）が含まれる。

また、複数の AI エージェントが連鎖して動作するマルチエージェント構成では、上位エージェントから下位エージェントへタスクが委譲される際に、上位エージェントが持つ全権限がそのまま引き継がれる設計になっていると、下位エージェントが本来不要な権限まで行使できてしまう（Confused Deputy 問題）。AI 提供者に対して、タスク委譲時に当該タスクの遂行に必要な最小限の権限のみを渡す設計（権限スコーピング）を求めることで、エージェント間の権限の連鎖的な拡大を防止する。

本要件は、連携先システムを持つ AI エージェントにおいて特に重要となる。連携先システムを持たない単体の生成 AI の場合は、重要度は相対的に低くなる。

（7） **参照データに対するアクセス制御と安全性検証**（T-7）

RAG を活用する場合、参照先の知識ベースに権限管理の不備があると、ユーザがアクセスすべきでない文書の内容が AI の回答に含まれてしまう。また、外部データにプロンプトインジェクションが仕込まれたり、悪意ある文書が混入したりすると、AI が意図しない動作をする。

AI 提供者に対して、以下のような機能の実装を求めることで、これらのリスクを低減できる。

- 参照データへのユーザ権限に応じたアクセス制御
- 外部データに含まれる悪意ある命令や不正コードの検知・遮断
- 外部データのサニタイズ（悪意ある部分の除去・変換）
- 参照データへのポイズニング検知・遮断

外部データに含まれる悪意ある命令や不正コードの検知・遮断については、関連機能の提供が始まっているが[49]、RAG が取り込む PDF・Web ページ・メール等の多様な形式の外部データに対するサニタイズは、実装が複雑であり、完全な対応は難しい。

参照データへのポイズニング検知については、RAG における知識ベースへの悪意あるデータ混入への対策として研究が進んでいる。ただし、特定の質問に対してのみ AI の回答を誘導する攻撃への防御機能

[48] Zero Trust for AI agents, https://claude.com/blog/zero-trust-for-ai-agents

[49] Microsoft, “Azure AI Content Safety のプロンプト シールド”, 2025 年 11 月 21 日更新, https://learn.microsoft.com/ja-jp/azure/ai-services/content-safety/concepts/jailbreak-detection, ユーザプロンプトや外部文書に含まれる敵対的入力攻撃を分析し、検知・遮断する機能として提供されている。

は、広く組み込まれるに至っていない[50]。

引き続き研究動向を注視し、AI 提供者が当該機能を提供する場合には、利用者側のデータ管理・承認運用と組み合わせて活用することが望ましい。

（8） **リソース消費の制御**（T-8）

攻撃者による意図的な大量リクエストや無限ループにより、システム停止（DoS）や API 利用料の異常な高騰（Economic Denial of Sustainability：EDoS）が発生し得る。

AI 提供者に対して、以下の機能実装を求めることで、可用性の維持とコストの暴走防止を実現する。特に AI エージェントは自律的に試行錯誤を繰り返すため、単体の生成 AI と比較してリソース消費の予測が困難であり、本要件の重要度が高い。

- リクエスト数・トークン数・処理時間等に対する最大消費量の上限での強制遮断機能（ハードリミット機能）
- 異常検知による自動停止機能

（9） **ログの取得**（T-9）

AI システムの安全な運用においては、利用状況の異常をリアルタイムに検知するとともに、インシデント発生時に「なぜその回答をしたのか」「どのデータを参照したのか」を事後的に検証できることが不可欠である[51]。特に AI エージェントは複数のステップを自律的に実行するため、各ステップの判断根拠が記録されていない場合、異常の早期発見や原因究明、再発防止が困難になる。

AI 提供者に対して、AI の推論プロセスや連携先システムの呼び出し履歴を含む詳細なログの記録・提供を求めることで、リアルタイムでの異常検知（リクエスト数の急増、通常と異なるデータアクセスパターン等）と、インシデント発生時の説明責任を果たせる体制の双方を確保する。なお、取得すべきログおよびその費用等については 6.3.2 節を参照されたい。

（10）**ハードウェアレベルの保護**（T-10）

マルチテナント環境では、ハードウェアの隙を狙った情報漏洩のリスクが存在する。特に、同じサーバに同居する他社の環境からメモリの処理状況などの物理的な挙動を盗み見されるサイバー攻撃（サイドチャネル攻撃）を受けたり、GPU メモリ内に残存する前処理データ（自社のプロンプトや社内データの一時的なキャッシュ）が流出したりする恐れがある。

AI 提供者に対して、以下のような機能の実装を求めることで、これらのリスクを低減できる。

- 処理後の GPU メモリの完全な初期化

---

[50] Wei Zou et al., “PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models”, 2024, https://arxiv.org/abs/2402.07867, RAG の知識ベースに少数の悪性テキストを混入させることで、特定の質問に対して攻撃者が選んだ回答を生成させ得ること、また評価された複数の防御策では十分に防げない場合があることが示されている。

[51] マルチエージェント環境においては、各 AI エージェントに固有の ID を振り当てる機能が、AI 開発者で整備されつつある。この機能により、実行したエージェントの ID と実行内容をログとして取得できる。

- 他社との利用環境の厳格な隔離
- ハードウェア層への最新セキュリティパッチの適用

なお、他社との利用環境の厳格な隔離は、専有ホスト・専用テナント等により実現可能な場合があり、規制対応や高機密データの処理において有効な選択肢となる。一方で、共有基盤と比較してコストが増加し、リソース効率が低下する場合があるほか、利用可能なリージョン、インスタンスタイプ、GPU 等が制限される可能性がある。そのため、リスク低減効果と費用・運用上の制約を踏まえて採用を判断し、AI 提供者に対して対応状況の確認・開示を求めることが望ましい。

ただし、いずれの対策も完全な防御を保証するものではなく、攻撃手法は継続的に進化しているため、AI 提供者の対応状況を定期的に確認することが望ましい。また、これらの対策は AI 提供者が採用するクラウド基盤の対応に依存する部分が大きく、AI 提供者に対してはその対応状況の確認・開示をあわせて求めることが望ましい。

ローカル LLM を採用する場合は、自社で GPU インフラを管理するため、マルチテナント特有のリスクは排除できるが、これらのハードウェアレベルの保護を自社の責任で実施する必要がある。

6.1.2 運用対策（O）

本項では、AI 管理者が自社の環境において実施すべき AI システムの設定・監視を中心とした対策を示す。6.1.1 節（T）で示した要件を AI 提供者と合意することを土台としつつ、自社の業務環境に即した対策を重ねることで、より実効的なセキュリティが確保される。自社の機密情報の定義、組織の ID 管理体制、ネットワーク構成、予算制約等は企業ごとに異なるため、AI 管理者がこれらの自社固有の事情に合わせて対策を設定・運用する必要がある。

（1） **入出力のフィルタリング設定**（O-1）

AI 提供者が用意したガードレールやフィルタリング機能に対し、自社固有のルールを設定する。

AI 提供者がオプトアウト機能（入力データをモデルの学習に利用しない設定）を提供している場合は、これを有効にする。これにより、自社の入力データがモデルの再学習に取り込まれ、他のユーザへの回答を通じて情報が漏洩するリスクを低減できる。

入力では、マイナンバー、クレジットカード番号、社員番号等の定型的な機密情報のパターンを定義し、AI への入力時に該当箇所のみを自動で検知・マスキング（匿名化）するよう設定する。例えば、顧客の個人情報を扱うコールセンター部門では、氏名・住所・電話番号のパターンをマスキングする対象として定義することで、オペレーターが AI に問い合わせ内容を入力しても個人情報が AI 提供者に送信されることを防止できる。

出力では、AI の回答に含まれてはならない情報やフレーズをブロックリストとして登録し、該当する語句が出力に含まれていた場合にその出力を遮断・削除する。

また、データの機密度（社外秘・極秘等）、人間によるレビュー状況（未確認・承認済等）、AI への入力可否（入力可・要匿名化・入力不可等）といった観点でデータラベルを付与して、ラベルの情報に基づいて AI への受け渡しを制御する。これにより、AI 提供者が提供する汎用的な防御機能を自社の情

報資産に合わせてカスタマイズし、自社特有の情報漏洩リスクを低減できる。

（2） **利用者のアクセス権限設定** （O-2）

自社の ID 管理基盤と AI システムを連携させ、「システムへの不正な侵入を防ぐアクセス制御（認証）」と「ログイン後の過剰なデータ閲覧や操作を防ぐ権限管理（認可）」をそれぞれ適切に一元管理する。

- アクセス制御
  AI システムへのアクセスには多要素認証（MFA）を必須として導入し、パスワード漏洩によるアカウント乗っ取りや不正利用を水際で防止する。また、API キー等の認証情報は専用シークレット管理ツールや定期的なキーローテーションで厳格に管理する。
- アクセス権限設定
  AI が参照できるデータ範囲をユーザの役職・部署に応じて制限し、AI エージェントが呼び出せるツールや実行できる操作を必要最小限に絞る。

例えば、社内の RAG システムにおいて、営業部門のユーザには顧客情報へのアクセスを許可する一方、製品開発情報へのアクセスは制限するといった部署ごとのアクセス制御を設定する。また、AI エージェントがコミュニケーションツール（チャット、メール等）と連携する場合は、「メッセージの閲覧のみ可能で、送信や削除は不可」といった操作権限の制限を設ける。これらの「認証」と「権限設定」を組み合わせた多層防御により、仮にユーザのアカウントが乗っ取られたり、AI が不正なプロンプトによる攻撃を受けたりした場合でも、被害の到達範囲を自社の権限設計に基づいて最小化できる。

（3） **リソース・コスト管理** （O-3）

AI の利用量に対して適切な上限を設定し、予期せぬコストの暴走やサービス停止を防止する。

1 リクエストあたりの最大トークン数（max tokens）にハードリミットを設定するほか、一定時間あたりの API リクエスト数（RPM：Requests Per Minute）およびトークン処理量（TPM：Tokens Per Minute）を制限し、月額予算リミットと組み合わせることで、過剰な API 利用によるコスト増大を抑制する。また、消費金額をリアルタイムで監視し、急激なコスト増加を検知した場合に自動停止や人間による承認制に切り替える仕組みを導入する。

例えば、AI エージェントが自律的に試行錯誤を繰り返す業務（市場調査の自動化等）では、月額の API 利用額が設定した閾値（例：50 万円）を超えた場合にアラートを発報して管理者の承認なしには追加実行できないようにする等の対策が考えられる。これにより、EDoS 攻撃や AI エージェントの処理暴走による想定外の巨額なコスト発生を防止できる。

（4） **インフラ・環境分離** （O-4）

AI システムを自社の社内ネットワーク内で他の重要システムから物理的・論理的に切り離し、万が一 AI システムが侵害された場合の被害範囲を限定する。

具体的には、AI システムの通信をプロキシ経由に限定してプライベート IP アドレス帯への直接アクセスを遮断するとともに、本番環境と検証環境を分離する。

例えば、AI システムのネットワークセグメントを社内の基幹システム（人事システム、会計システム等）から分離し、AI システムから基幹システムへの通信は特定の API エンドポイントのみに限定する。これにより、AI システムが攻撃者に掌握されたとしても、基幹システムへの横展開を防止できる。

（5） **ログの有効化**（O-5）

AI システムの利用状況をリアルタイムで監視・検知し、かつ事後的に追跡可能とするため、AI システムの全操作ログ（いつ・誰が・どのサービスに・何を送ったか）および AI システムの推論プロセス（思考の過程やツール呼び出し履歴）を記録できる状態に設定する。これにより、不正アクセスや不審な挙動のリアルタイムな検知・遮断を可能にするとともに、情報漏洩や権限外操作などのインシデントが発生した際にも、このログが確実に取得されていることが、原因究明や被害範囲の特定、および経営層や外部への説明責任を果たすための証跡となる。

なお、ここで取得したログをどのように監視・分析するか（SIEM[52]連携や不審なアクセスパターンの検知等）に対する具体的な手法や、実務における効率的なログ監査の運用方法については、6.3.2 節「ログ監査の実務設計」にて詳述しているため、システム設計および運用体制の構築にあたってはそちらを参照されたい。

6.1.3 人的統制（H）

本項では、6.1.1 節（T）および 6.1.2 節（O）の技術要件・運用対策ではカバーしきれないリスクに対して、AI システムを利用する「人」の判断や特定の AI システム特有のルール構築を中心とした対策を示す。

前項までの技術的・運用的な防衛層を土台としつつ、最終的に AI システムに指示を出しその出力を実業務に適用する人間のレイヤーに対策を重ねることで、多層防御としての実効性が確保される。AI システムが自律的に外部操作を実行する際のリスク許容度や、部門ごとの RAG システムが参照する情報の性質は個別の環境ごとに異なる。そのため、重大な操作の前に人の判断を挟む「人間介入・承認プロセス」の実装や、現場の利用実態に即した「個別システム特有のルール策定」を、自社のガバナンス事情に合わせて柔軟に選択・運用する必要がある。

（1） **人間介入・承認プロセス**（H-1）

AI システムの出力を業務判断や外部送信に利用する前に、必ず人間の確認・承認を経る業務フローを整備する。具体的には、以下のような高リスクな操作については、AI システムが自動実行する前に人間の承認を必須とするワークフローを組み込むべきである。

- 決済の実行
- データの外部送信
- システム設定の変更

---

[52] SIEM（Security Information and Event Management）とは、システムやセキュリティ機器等からログやイベント情報を収集し、単一の画面で分析・監視できるようにする仕組みまたは製品のこと。異常の検知やインシデント調査に用いられる。

- 顧客への回答送信

留意点として、承認が形骸化しないよう、承認対象を重要度に応じて絞り込むことが重要である。全ての出力に承認を求めると承認疲れが起き、形だけの確認に陥るリスクがある。人間の承認が必要な操作を業務リスクに基づいて明確に定義し、それ以外は AI システムの自律実行を許容するという判断軸を設けることが実効性の鍵となる。

（2） **個別システム特有のルール策定**（H-2）

全社 AI 利活用ガイドライン（第 2 章）はすべての AI 利用に共通するルールを定めるものであるが、個別の AI システムでは、その利用目的や連携先、扱うデータの特性に応じて、追加のルールが必要となる場合がある。

例えば、以下のようなルールが考えられる。

- AI が参照する社内データの追加・更新は、データの所管部門の承認を得た上で実施する。
- AI エージェントが連携する外部ツールの追加・変更は、セキュリティ部門の承認を必須とする。
- AI の出力を顧客向けに利用する場合は、出力内容の事実確認と権利侵害チェックを経た上で利用する。
- 特定の業務領域（法務判断、財務報告等）では、AI の出力を最終判断の根拠として単独で使用しない。
- AI システムが扱うデータの機密性に応じて、利用可能な時間帯や場所（社内ネットワーク限定等）を制限する。

これらのルールは、第 5 章で評価した固有リスクの高さに応じて策定する。固有リスクが高いシステムほど、より詳細かつ厳格なルールが必要となる。一方、固有リスクが低いシステムに対して過度なルールを設けた場合は、AI 利用者の利便性を損ない、かえってルールの形骸化を招くリスクがある。

留意点として、個別ルールは全社 AI 利活用ガイドラインと矛盾しないよう整合を取る必要がある。また、AI システムの利用目的や連携先が変更された場合は、ルールの見直しもあわせて実施する。ルールの策定にあたっては、AI 利用者、セキュリティ部門、および必要に応じて法務部門と連携することが望ましい。また、複数部門に承認プロセスを設定する場合は、運用開始前に AI 連絡会議等で対象部署と合意を得る必要がある。

6.2 発生被害と対策の関係性

本節では、第 4 章で定義した 7 つの発生被害（A～G）と 6.1 節で示した対策との対応関係をマトリクス表として整理する。すべての対策を一律に実施する必要はなく、第 5 章で評価した対象システムの固有リスクの高さに応じて、優先度の高い対策から段階的に取り組むことが現実的である。各発生被害に対する対策の具体的な組み合わせ方、優先順位、残存リスク、導入・運用上の留意点については、付録「発生被害ごとの対策の進め方」に詳述している。付録では、各発生被害について「まず取り組むべき基本対策」と「固有リスクに応じて追加する対策」の 2 段階に分けて整理しているため、対象システムの固有リスクの高さや自社のリソース制約に応じた対策の選択が可能である。

本節の活用方法は以下のとおりである。

① 第 5 章のリスク評価結果をもとに、自社の AI システムにおいて対応が必要な発生被害とその固有リスクの高さを確認する。
② 表 6-1 で該当する発生被害（A～G）を確認し、●が付いている対策を把握する。
③ 付録「発生被害ごとの対策の進め方」で当該発生被害の項を参照し、対策の組み合わせ方、優先順位、残存リスク、導入・運用上の留意点を確認する。

以下の表 6-1 は、6.1 節の各要件・対策が、どの発生被害のリスク低減に有効であるかを示す。●は当該発生被害に対して有効な対策であることを意味する。

表 6-1　発生被害と対策の対応

要件項目 / 対策カテゴリ	A	B	C	D	E	F	G
【6.1.1】技術要件（T）							
(1) 学習データの機密性保持（T-1）	●						
(2) 指示とデータの構造的分離（T-2）	●	●				●	
(3) 真正性・透明性の確保（T-3）	●				●	●	
(4) 入出力の監視とフィルタリング機能（T-4）	●	●	●		●	●	
(5) ユーザ環境分離とデータ処理環境の確保（T-5）	●	●					
(6) 連携先システムにおける実行権限の最小化（T-6）	●	●	●			●	
(7) 参照データに対するアクセス制御と安全性検証（T-7）	●		●				
(8) リソース消費の制御（T-8）				●			●
(9) ログの取得（T-9）	●	●	●	●	●	●	●
(10) ハードウェアレベルの保護（T-10）	●						
【6.1.2】運用対策（O）							
(1) 入出力のフィルタリング設定（O-1）	●	●	●		●	●	
(2) 利用者のアクセス権限設定（O-2）	●	●	●			●	●
(3) リソース・コスト管理（O-3）				●			●
(4) インフラ・環境分離（O-4）	●	●		●		●	
(5) ログの有効化（O-5）	●	●	●	●	●	●	●
【6.1.3】人的統制（H）							
(1) 人間介入・承認プロセス（H-1）		●	●		●	●	
(2) 個別システム特有のルール策定（H-2）	●	●	●		●		

A:情報の漏洩　B:付与権限外の操作　C:システムや内部データの改ざん　D:AI システムのサービス停止

E:虚偽や低品質な出力の利用および出力の誤用　F:第三者への加害　G:過剰課金の発生

6.3 AI 運用における実務課題と対策のアプローチ

6.1 節では対策の一覧を、6.2 節では発生被害と対策の対応関係を、付録では各発生被害に対する対策の具体的な進め方を示した。しかし、これらの対策を実際に運用しようとすると、多くの企業が共通して直面する実務上の課題がある。

本節では、特に対策の優先度が高い 2 つの実務課題（情報入力統制およびログ監査の実務設計）について、現実的な対応アプローチを示す。いずれの課題も、完璧な対応を初めから目指すのではなく、自社のリスクとリソースに応じて段階的に導入することが重要である。

6.3.1 情報入力統制

第 2 章で整備した全社 AI 利活用ガイドラインにより、組織承認 AI の利用やデータ入力基準を定め、周知していたとしても、従業員が許可されたレベルを超える情報を AI に入力してしまうリスクがある。例えば、「業界最大手 A 社を買収する件についてのプレスリリース案を書いて」といったプロンプトには、文脈の中に重大な機密情報が含まれている。しかし、このような文脈に埋め込まれた機密情報を技術的に完全に検知することは極めて困難である。

本課題は、運用ルールや利用画面上での注意喚起により利用者の入力行動を統制しつつ、リスクや業務影響に応じて、技術的に検知・制御する範囲を段階的に拡大していくという実務上の判断が求められるものである。

■対策のアプローチ

Step 1：利用画面上での入力ルール提示

第 2 章の全社 AI 利活用ガイドラインに基づき、個別の AI システムごとに「入力してよい情報の範囲」を具体的に定める（H-2 個別システム特有のルール策定）。例えば、「機密レベル○○以上の情報は AI に入力禁止」等のルールを利用画面上に常時表示する。加えて、AI システムのログイン時にポップアップ警告で表示することで、心理的に入力しにくい状態をつくる。導入コストは最小限であり、情報入力統制の中核的な対策となる。

Step 2：パターン検知によるマスキング

O-1「入出力のフィルタリング」で示したマスキング機能を活用し、定型的な機密情報パターン（マイナンバー、クレジットカード番号、社員番号等）を自動検知・マスキングする。過度に厳格な検知ルールは、本来入力可能な情報までマスキングし、AI の回答品質低下、再入力や例外申請の増加、業務遅延を招く可能性がある。そのため、初期段階では誤検知が少なく業務影響を把握しやすい情報種別から開始し、検知精度と業務影響のバランスを確認しながら、対象を段階的に拡大する。

Step 3：DLP や AI ゲートウェイの導入

機密情報を扱う部門や高リスク AI システムでは、DLP や AI ゲートウェイを活用し、AI システムへの入力制御を強化する。DLP により、端末やブラウザ上での機密情報のコピー・貼り付けや入力を検知・制御できる。また、社内ネットワークと AI システムの間に AI ゲートウェイを設置することで、AI システムへ送信さ

れる前の入力内容を検査し、機密情報の検知、マスキング、送信ブロック等を集中管理できる。

ただし、これらの導入にはネットワークや端末設定の変更、業務影響調査、例外運用の設計が必要となり、コストと人的リソースを要する。そのため、全社一斉導入ではなく、まずは機密情報を扱う部門、高リスク AI システム、または明らかに危険な入力・送信経路に対象を絞って適用し、効果と業務影響を確認しながら適用範囲を拡大することが現実的である。

■残存リスク

- 文脈に含まれる機密情報は、パターン検知では検出が極めて困難である。
- 画像、音声、PDF 等の非構造化データに含まれる機密情報については、現時点では検知精度や運用負荷に課題があり、検出漏れが生じうる。

■6.1 節・付録との関連

本課題は、T-4「入出力の監視とフィルタリング機能」、O-1「入出力のフィルタリング」、H-2「個別システム特有のルール策定」、および付録「A. 情報の漏洩」の対策 1・4 と直接関連する。また、全社的なデータ入力基準やセキュリティ教育については第 2 章を参照されたい。

6.3.2 ログ監査の実務設計

T-9「ログの取得」および O-5「ログの有効化」で示したとおり、ログの記録は情報漏洩や権限外操作等の検知・追跡に不可欠である。しかし、従業員が AI システムを利用するたびに発生する膨大なログ（プロンプトや出力結果）を、AI 管理者やセキュリティ担当者の限られたリソースで効率的かつ実効性を持って監査することは容易ではない。全件の目視チェックは物理的に不可能であり、コストとリスクを見極めた現実的な運用設計が求められる。

■取得すべきログの要件

インシデント発生時の原因究明や不正利用の検知を目的として、以下のログを取得する。

- 利用者・端末：ユーザ ID、所属部門、端末情報（MAC アドレス、PC 名）、IP アドレス
- 利用時間：アクセスおよびリクエスト実行の日時
- 利用システム・サービス：どの AI システム、どの基盤モデルを利用したか
- 入力：入力したプロンプトの生データ、添付ファイル名
- 出力：AI が生成した出力内容、消費したトークン数
- 制御：DLP や AI ゲートウェイによるブロック履歴、マスキングの実行履歴

保存期間については、すべての入出力テキストを長期保存するとストレージ費用が膨大になる。日時やユーザ ID 等の「メタデータ」は長期保存（例：1 年）とし、容量を要する「プロンプトや出力テキスト（ペイロード）」は短期保存（例：3 ヶ月）とするなど、データの性質に応じたライフサイクル管理を行う。

■対策のアプローチ

ログ監査の仕組みは、組織の成熟度や要求されるセキュリティレベルに応じて段階的に高度化していくこ

とが望ましい。ここでは以下の 3 段階でのアプローチを提案する。なお、この成熟度に応じた段階的なアプローチは、「Zero Trust for AI agents[53]」における階層モデル（Foundation / Enterprise / Advanced）とも整合するものである。

レベル 1 ：ログの蓄積と事後確認（Foundation 相当）

リアルタイム監視は行わず、ログを蓄積のみ行う。インシデント発生時のフォレンジック（事後調査）や月次等の定期確認に使用する。ただし、重要システムではごく一部のログだけでも、毎日、人が確認することが望ましい。ログ蓄積のみの場合、導入ハードルは最も低いが、未然防止・早期発見はできない。AI システムを導入するすべての企業が最低限実施すべきレベルである。なお、ここで確立したベースラインは、異常検知の基準になるだけでなく、侵害後に正常な状態へ戻すための復旧基準となる。

レベル 2 ：キーワード・パターン検知による自動通知（Enterprise 相当）

あらかじめ登録したキーワードやパターン（機密情報の名称、禁止操作のコマンド等）を検知し、自動で担当者に通知する。担当者が通知を確認し、問題があればサービスの遮断や当該従業員への連絡を行う。一定の費用はかかるが、明らかなルール違反を自動で検知できる。多くの SIEM 製品やログ管理ツールで実現可能である。なお、確立したベースラインと現在の挙動を比較し、逸脱を検知するが、急激な異常だけでなく、メモリ汚染などによる緩やかな挙動変化も対象にすることが重要である。

レベル 3 ：監査 AI システムの導入（Advanced 相当）

AI システムへの入力を監査用の AI システムにも読ませ、キーワード検知以上の文脈理解に基づく判定を行う。ただし、AI システムを使うたびに監査 AI システムも使用するため利用料が実質 2 倍になるほか、監査 AI システム自体の統制・精度チューニング・監査 AI システムの監査も必要となる。費用対効果を慎重に見極める必要がある。なお、異常を検知した場合、被害を抑えるための対応を行うが、全ての判断を自動化するのではなく、重要な判断は人が行う必要がある。

■残存リスク

- ログ自体に機密情報が含まれる場合、ログの閲覧権限の管理が必要となる（部門によってはログを閲覧できない可能性がある）。
- いずれのレベルでも検知漏れは発生し得るため、ログ監査は他の対策（入出力フィルタリング、アクセス権限設定等）との多層防御の一部として位置づける。

■6.1・付録との関連

本課題は、T-9「ログの取得」、O-5「ログの有効化」、および付録の各発生被害における「ログの有効化」の対策と直接関連する。

[53] Anthropic 社が 2026 年 5 月に提唱したエンタープライズ向けのセキュリティフレームワーク, Anthropic, Zero Trust For AI agents ,https://claude.com/blog/zero-trust-for-ai-agents

第 7 章　おわりに

生成 AI や AI エージェントの進化は、我々の想像を超える速度で進んでいる。今や企業の業務に組み込まれはじめ、AI の活用が企業競争力を左右する時代となっている。AI を積極的に取り入れる企業と、様子見を続ける企業との差は、着実に広がりつつあり、AI の活用はもはや企業存続のための必須条件になりつつあると言っても過言ではない。

将来を見渡せば、各社が自社の業務ノウハウや知見を学習させた独自の AI システムを保有・運用する時代が、さほど遠くない将来に訪れるかもしれない。そのとき、AI をいかに安全かつ効果的に扱うかという能力が、企業の競争優位の一要素を成すことになるだろう。

しかしながら、こうした機運が高まる一方で、AI の安全な活用に関するガイドラインは国内外で乱立しており、民間企業の担当者が「結局、何をしたらいいのか。」と迷ってしまう実情があるだろう。国際機関、各国政府、業界団体、研究機関がそれぞれの観点からガイドラインを発行しており、その数は膨大である。

こうした問題意識のもと、民間企業が生成 AI および AI エージェントを業務活用するうえで何に気をつけるべきかという問いをメンバー間で議論し、その成果を本手引書としてまとめた。

まず、全社 AI ガバナンスの実務的な指針を示した。CAIO 設置による責任体制の明確化から、全社 AI 利活用ガイドラインの策定、シャドーAI への対応方針に至るまで、組織が AI を統制するために必要な施策を体系化した。ガイドラインを一度策定して終わりにするのではなく、現状の把握を行い、技術進化に応じて継続的に更新する「継続的な適応」の重要性を示した点も、本書の主張の一つである。

次に、AI システムのライフサイクル全体を通じたセキュリティ管理の枠組みを整理した。企画・調達・技術検証・運用前準備・運用・廃棄の各フェーズで実施すべき事項を示した。

また AI システムがもたらしうるリスクを、ユーザ企業が実際に被る業務上の実害という観点から 7 つの発生被害として体系化し、発生被害ごとにリスクを評価する実践的な手法を提示した。

さらに、セキュリティ対策を「技術要件」「運用対策」「人的統制」の 3 つに分類し、実施主体を明確にした枠組みを提示した。予算・リソースの制約が大きい企業においても、残存リスクや留意点を示すことで、具体的に動けるよう設計している。

本手引書が AI セキュリティのすべてを網羅したものではない。また、技術の進化とともに、新たなリスクや論点は今後も次々と生まれ続けるだろう。それでも、民間企業において AI 活用を推進する立場の方々や、セキュリティを担う方々にとって、本手引書が新たな気づきのきっかけとなり、自社における取組みを一歩前に進めるための一助となれば幸いである。

謝辞

本書の作成にあたりまして、ご協力いただきました企業の皆様には多大なるご支援・ご尽力を賜りました。この場を借りて心より御礼申し上げます。

また、産業サイバーセキュリティセンター 中核人材育成プログラムの講師である、門林 雄基先生、小林 和真先生には、本書の元となるプロジェクトのメンターとして、ご指導・ご助言を賜りました。改めて御礼申し上げます。

そして、本書の作成や本プロジェクトをともに実施した、下記メンバーの皆様にも感謝を伝えたいと思います。

セキュリティ担当者のための生成 AI セキュリティプロジェクト

ドキュメントチーム　メンバー

【プロジェクトリーダー】

星　凜太郎　　山田　敏大

【チームリーダー】

和田　浩樹

【執筆メンバー】

亀井　祐樹　　佐藤　舞　　中島　しおり

【レビューメンバー】

坂脇　孝弘　　島袋　洋一　　下出　有実

榛葉　光　　向井　耕司

付録　発生被害ごとの対策の進め方

発生被害 A：情報の漏洩

情報の漏洩に対しては、表 6-1 のとおり多くの対策が有効である。ただし、これらの対策を全て同時に導入する必要はなく、自社のリスクと状況に応じて複数の対策を組み合わせて実施することが重要である。

■まず取り組むべき基本対策

以下の対策 1 から対策 4 は、AI を業務に導入するすべての企業が、導入初期の段階で取り組むべき対策である。いずれも大規模な投資を必要とせず、既存の業務プロセスの延長で実施可能なものが中心である。

（1） 入出力のフィルタリング

- **組み合わせる対策**
  - 技術要件：入出力の監視とフィルタリング機能（T-4）
  - 運用対策：入出力のフィルタリング設定（O-1）
- **なぜこの組み合わせが有効か**

  T-4 で AI 提供者に汎用ガードレール・マスキング機能の基盤を求め、O-1 で自社固有の機密情報パターンを設定することで、汎用的な検知だけではなく、自社特有の情報も保護できる。また、入力側と出力側の両方でフィルタリングが機能するため、二重の防御が実現できる。
- **残存リスク**
  - ブロックリストに未登録のパターンは検出漏れが生じる。
  - 機密情報が言い換えや要約された形で出力された場合、ルールベースの検知は困難である上、動的マスキング機能を用いても、すべての個人情報や機密情報の漏洩を防ぐことはできない。
- **導入・運用上の留意点**
  - 汎用ガードレール・マスキング機能の有無と仕様を確認した上で、自社固有の機密情報パターン設定を要求する。設定が困難な場合は、AI Gateway や AI DLP を組み合わせる方法を検討する。
  - マスキング対象の情報パターンの洗い出しには、セキュリティに関する知識と業務知識の両方が必要であり、チューニングのための設定工数が必要である。また、ブロックリストの継続的更新には担当者の明確な割り当てと更新サイクルの制度化が必要である。
  - 入力フィルタリングは AI モデルへのデータ送信を防ぐ効果を有するが、同一システム内での他ユーザへの出力漏洩を防ぐには、別途アクセス権限設定（対策 2）による対応が必要となる。

（2） 連携先システムへのアクセス制御

- **組み合わせる対策**
  - 技術要件：連携先システムにおける実行権限の最小化（T-6）
  - 運用対策：利用者のアクセス権限設定（O-2）
- **なぜこの組み合わせが有効か**

T-6 で AI 提供者に連携先システムへのアクセス権限を必要最小限に制限することを求め、O-2 で利用者ごとに AI システム経由でアクセスできるデータ範囲を制御することで、「AI が到達できるデータ範囲」と「利用者が AI 経由で引き出せるデータ範囲」の両方を制限できる。どちらか一方だけでは、もう一方の権限設定の不備を補完できないため、両面からの制限が重要である。

- **残存リスク**
  - 権限設定の不備・更新漏れ（異動者・退職者の権限変更・無効化漏れ等）によって、不要なデータにアクセス可能なリスクが残る。
  - プロンプトインジェクション攻撃により、正規ユーザの権限範囲内であっても本来意図しないデータを AI から引き出させる攻撃は対応できない。
- **導入・運用上の留意点**
  - 既に社内で運用している ID 管理の仕組みを AI システムにも適用することで、実現可能な場合も多い。アクセス権限の設計は人事部門（異動情報）や各事業部門（業務上必要なアクセス範囲）との連携が不可欠である。

(3) ログの有効化による漏洩監視

- **組み合わせる対策**
  - 技術要件：ログの取得（T-9）
  - 運用対策：ログの有効化（O-5）
- **なぜこの組み合わせが有効か**

T-9 で AI 提供者に求めた詳細なログ記録基盤の上に、O-5 で自社の SIEM との連携や監視ルールを設定することで、漏洩の検知・追跡の体制を構築できる。これにより、インシデントが発生した場合に原因究明と再発防止を迅速に行えるほか、ログの存在自体が内部不正の抑止力となる。

- **残存リスク**
  - ログ分析体制が未整備であれば、異常検知が遅延する。AI 提供者によっては、仕様により、ログの粒度が限定される場合がある。
  - ログに機密情報が含まれる場合、ログの保管場所や権限設定が不適切だと、ログ自体が新たな漏洩経路になる。
- **導入・運用上の留意点**

- 本対策は、事後検知や追跡が目的であるため、漏洩の発生を防ぐことはできない。他対策と組み合わせる必要がある。
- まずはログを記録することから始め、SIEM 連携やリアルタイム分析は段階的に整備する。SIEM 連携やリアルタイム分析の実施には、ネットワークやセキュリティに関する専門知識が必要である。
- 監視ルールの設計はセキュリティ部門、アラート発報時の対応フローは情報システム部門、経営層との事前合意が必要である。

(4) 個別システム特有のルール策定

- **対象の対策**
  - 人的統制：個別システム特有のルール策定（H-2）
- **なぜこの対策が有効か**

  全社 AI 利活用ガイドライン（第 2 章）に加え、当該 AI システムが扱うデータの機密性や利用形態に応じた個別ルールを策定することで、リソースの影響で技術的対策を実施できない場合や、技術的対策では防ぎきれない人的要因による情報の漏洩に有効である。ルールの存在自体が利用者への心理的抑止力として機能するとともに、違反が発生した際の責任の所在と対応の根拠を明確にする効果もある。
- **残存リスク**
  - 周知・教育が不十分な場合や、業務実態とルールが乖離している場合には、ルールが形骸化する恐れがある。
  - ルールの順守状況を技術的に検証する手段がない場合は、ルール違反を見逃すリスクが残る。
- **導入・運用上の留意点**
  - ルールは AI システムの利用部門とセキュリティ部門、また必要に応じ法務部門が連携して策定し、業務実態と乖離しないようにする。
  - 策定したルールは AI システムのログイン画面や利用画面上に常時表示することで、ルールをすぐに閲覧できるようにするとともに、心理的抑止効果を高めることができる。

■固有リスクに応じて追加する対策

以下は、第 5 章の評価で情報漏洩の固有リスクが「中」以上と判定された場合や特定の利用形態に該当する場合に追加で検討する対策である。

(5) ネットワーク分離と環境隔離

- **組み合わせる対策**
  - 技術要件：ユーザ環境分離とデータ処理環境の確保（T-5）
  - 技術要件：ハードウェアレベルの保護（T-10）

- 運用対策：インフラ・環境分離（O-4）

- **どのような場合に有効か**

AI システムが基幹システム等の重要なシステムと同一ネットワーク上にある場合や、RAG 等で扱うデータの機密性が特に高い場合に有効である。

- **なぜこの組み合わせが有効か**

情報漏洩の経路は「AI 提供者側」と「自社ネットワーク内」の二方向に存在する。AI 提供者側に対しては T-5 によるテナント間のソフトウェアレベルの隔離と T-10 による GPU メモリ初期化・サイドチャネル攻撃対策のハードウェアレベルの隔離を組み合わせて求めることで、外部環境での二層の防御を実現できる。これに自社ネットワーク内で AI システムを重要システムから切り離す O-4 を加えることで、外部・内部の両方の漏洩経路を多層的に遮断できる。

なお、機密性が特に高いデータを扱う場合は、ローカル LLM の活用によりデータが自社外に一切出ないアーキテクチャも検討する。

- **残存リスク**

  - 内部の正規ユーザによる意図的な情報持ち出しは防げない。
  - ローカル LLM を採用する場合でも自社環境内の設定不備による情報漏洩リスクは残る上、LLM のオープンソースモデルのサプライチェーンリスク（バックドアや悪意のある学習データ混入）についても考慮が必要である。
  - ハードウェアレベルの未知の脆弱性（ゼロデイ）への対応には限界があり、攻撃手法は継続的に進化しているため、AI 提供者の対応状況を定期的に確認することが望ましい。

- **導入・運用上の留意点**

  - ネットワーク構成の変更は情報システム部門との調整が必要であり、工数とコストが発生する。
  - ローカル LLM は GPU 等のインフラコストに加え、オープンソースモデルの安全なデプロイやセキュアなインフラ保守、継続的なモデル管理を自社で主体的に遂行できる専門人材が必要であり、データの機密性とコスト、体制のバランスで判断する。
  - O-4 の実装にあたっては、AI システムから外部への意図しないデータ送信を遮断する外向き通信制御を設定することも有効である。
  - データ保存場所の要件は、法務部門や情報システム部門との事前協議が必要である。

(6) RAG の参照データに対するアクセス制御

- **対象の対策**

  - 技術要件：参照データに対するアクセス制御と安全性検証（T-7）

- **どのような場合に有効か**

RAG を活用して社内文書を AI に参照させる場合に有効である。特に部署や役職によってアクセスできる文書が異なる場合や、機密性の異なる文書が混在する知識ベースを運用する場合は重要度が高い。

- **なぜこの対策が有効か**

RAG のアクセス制御が不十分な場合、本来その利用者がアクセスすべきでない文書の内容が AI システムの回答に混入するリスクがある。T-7 により利用者の権限に応じて参照可能な文書の範囲を AI 提供者側で制御することでこのリスクを低減できる。なお、O-1 のデータラベルが AI への入出力時の制御を担うのに対し、T-7 は RAG 検索時の参照範囲を制御するものであり、対策 2（O-2）の利用者認証・認可とあわせて多層的な防御を構成する。

- **残存リスク**
  - RAG はベクトル類似性に基づいて文書を検索するため、アクセス権限外の文書の断片が類似文書として検索結果に混入するリスクが完全には排除できない。
  - プロンプトインジェクション攻撃によりアクセス制御のロジックをバイパスされる可能性がある。（この対策は後述の対策 7 を参照されたい。）
  - 権限ラベルの付与漏れや更新漏れにより、意図しないアクセスが発生するリスクが残る。
- **導入・運用上の留意点**
  - 本対策の実現手段としては、知識ベース内の各文書に部署、役職、機密性等のアクセス権限ラベルを付与し、RAG 検索時に利用者の権限と照合（対策 2 参照）して参照可能な文書を絞り込む方法がある。ただしチャンク単位での細粒度アクセス制御は技術的難易度が高く、多くの RAG 製品で対応が未成熟である。AI 提供者の本機能提供可否は、データ所管部門やセキュリティ部門と連携して、AI 提供者に調達段階で確認すべきである。

（7） プロンプトインジェクション対策の強化

- **対象の対策**
  - 技術要件：指示とデータの構造的分離（T-2）
- **どのような場合に有効か**

AI システムがユーザからのアップロードファイルや Web サイト等の外部データを参照する場合、AI エージェントが連携先システムと接続する場合に有効である。

- **なぜこの組み合わせが有効か**

プロンプトインジェクションは、AI がシステムの指示とユーザ入力・外部データを内部的に区別できないことに起因するため、システム指示と入力データを構造的に分離するアーキテクチャの導入は有効である。

- **残存リスク**

- 未知の攻撃手法や巧妙に偽装された指示に対しては完全な防御を保証できない。
- AI 提供者の更新が、新たな攻撃手法の出現に追いつかない期間が生じる可能性がある。

- **導入・運用上の留意点**
  - 現時点では多くの AI モデルがこの分離を十分に実現できておらず、対応可能な AI 提供者が限定される。本対策に過度に依存せず、対策 1～4 との多層防御の一部として位置づける必要がある。
  - 調達仕様書に要件として盛り込み、AI 提供者の安全性テスト報告書の開示を求める。対応状況は AI 提供者の選定段階で確認する（第 3 章参照）。
  - 外部データを参照する構成では、参照先データソースの信頼性を事前に評価し、不特定多数が書き込める外部ソースの参照は最小限にとどめることも有効な補完策となる。

(8) モデルの透明性確保

- **対象の対策**
  - 技術要件：真正性・透明性の確保（T-3）
- **どのような場合に有効か**

  AI の利用規模が大きい場合やコンプライアンス要件で AI 利用に関する説明責任が求められる場合に有効である。
- **なぜこの対策が有効か**

  AI 提供者に SBOM やモデル構成情報の開示を求めることで、モデルの構成部品に起因する漏洩リスクを可視化し、脆弱性が発見された際に迅速な対応判断を可能にする。
- **残存リスク**
  - SBOM を取得しても脆弱性情報との継続的な突合・精査体制が整っていない場合、取得するだけで対処が追いつかず形骸化するリスクがある。
- **導入・運用上の留意点**
  - 本対策は、リスクの可視化と早期発見が目的であるため、漏洩の発生を防ぐことはできない。他対策と組み合わせる必要がある。
  - 商用利用可能なモデルでは SBOM やモデル構成情報の開示に応じないケースが大半であり、本対策を実現できる場面は限られる。SBOM 提出を調達要件に含めるところから始めるとよい。
  - SBOM の継続的な脆弱性精査のプロセスについては第 3 章のライフサイクル管理で説明する。精査には AI モデルのサプライチェーン構造と脆弱性データベースの活用に関する専門知識が必要である。

(9) 学習データとハードウェアレベルの対策

- **組み合わせる対策**
  - 技術要件：学習データの機密性保持（T-1）
  - 技術要件：ハードウェアレベルの保護（T-10）
- **どのような場合に有効か**

  自社の機密性の高い業務データを RAG の参照データとして登録する場合や、自社専用環境にローカル LLM を構築して機密性の高いデータを扱う場合に有効である。
- **なぜこの組み合わせが有効か**

  LLM は学習データに含まれる情報を推論時に再現してしまう場合がある。T-1 により AI 提供者に対して差分プライバシー等の技術的手法が適用されていることの確認・開示を求めることで、学習データ経由の情報漏洩リスクを低減できる。
- **残存リスク**
  - 差分プライバシーはプライバシー保護強度を高めるほどモデルの出力精度が低下するトレードオフがある。（6.1.1 節（1）説明参照）
- **導入・運用上の留意点**
  - AI 提供者に対して、調達フェーズで本対策の適用状況を確認する必要がある。
  - 精度と安全性のトレードオフについて AI 提供者と事前協議が必要である。

■対策の進め方のまとめ

本発生被害への対策は、以下のステップで段階的に進めることを推奨する。

① AI 提供者のサービスに備わる入出力のフィルタリング機能・アクセス制御・ログ機能を有効化し、個別ルールを策定する（対策 1～4）。

② 第 5 章で評価した固有リスクの高さに応じて、ネットワーク分離、RAG のアクセス制御、プロンプトインジェクション対策等を追加する（対策 5～9）。

①は追加コストが比較的小さく、AI を導入するすべての企業が実施可能な対策である。②は自社の固有リスクや利用形態に応じて、優先度の高いものから段階的に取り組む。対策 7～9 は AI 提供者の対応状況に依存するため実現が困難なケースも多く、対応可能な AI 提供者の選定基準として活用することが現実的である。

発生被害 B：付与権限外の操作

付与権限外の操作は、特に AI エージェントが連携先システムやデータベースに接続する構成で特にリスクが高まる。単体の生成 AI の場合は連携先システムがないため、本被害のリスクは相対的に低い。

■まず取り組むべき基本対策

（1） 実行権限の最小化とツール制御

- **組み合わせる対策**

- 技術要件：連携先システムにおける実行権限の最小化（T-6）
- 運用対策：利用者のアクセス権限設定（O-2）

- **なぜこの組み合わせが有効か**

T-6 で AI システムの操作を最小限に制限し、O-2 で利用者ごとに AI 経由でアクセスできるシステムとデータの範囲を制御することで、AI システム側・利用者側の両方から操作範囲を絞ることができる。

- **残存リスク**

- 正規権限の範囲内での悪用は防げない。
- ツール単位・引数単位での細粒度な権限制御は AI 提供者の実装に依存するため、十分な粒度での制御が困難なケースがある。

- **導入・運用上の留意点**

- MCP サーバ等のツール定義を利用する場合は、呼び出せる機能・引数・入力値の範囲を厳密に制限する。実装には AI に関する専門知識が必要である。

(2) 人間介入・承認プロセスの整備

- **対象の対策**

- 人的統制：人間介入・承認プロセス（H-1）

- **なぜこの対策が有効か**

高リスクな操作（データの削除・更新、外部への送信、システム設定の変更等）の実行前に人間の承認を求めることで、技術的対策をすり抜けた攻撃に対する最終防衛線となる。

- **残存リスク**

- 承認対象が多い場合、承認疲れにより形骸化する。

- **導入・運用上の留意点**

- 人間の承認が必要な操作を業務リスクに基づいて明確に定義し、低リスクの操作は AI の自律実行を許容することで、実効性を高めることができる。
- 承認フローの設計は実業務に合わせて AI 利用部門との連携が必要である。

(3) 入出力のフィルタリングとログ取得

- **組み合わせる対策**

- 技術要件：入出力の監視とフィルタリング機能（T-4）
- 運用対策：入出力のフィルタリング設定（O-1）
- 技術要件：ログの取得（T-9）
- 運用対策：ログの有効化（O-5）

- **なぜこの組み合わせが有効か**

T-4/O-1 による入出力フィルタリングと、T-9/O-5 によるフィルタリングログにより、「どのツールが、どの引数で、何回呼び出されたか」を記録することで、異常操作の検知と事後の追跡体制を構築できる。

- **残存リスク**
  - 未知の攻撃パターンはフィルタリングで検知できない。
  - AI 提供者によってはツール呼び出し単位の詳細なログが取得できないケースがある。
  - ログ分析体制が不十分な場合は、異常検知への対応が遅延する。
- **導入・運用上の留意点**
  - ブロックリストの継続的な更新が必要である。
  - ログはまず記録することから始め、分析体制は段階的に整備する。

■固有リスクに応じて追加する対策

以下は、第 5 章の評価で付与権限外の固有リスクが「中」以上と判定された場合や AI エージェントが重要なシステムと連携する場合に追加で検討する対策である。

（4） プロンプトインジェクション対策の強化

- **対象の対策**
  - 技術要件：指示とデータの構造的分離（T-2）
- **どのような場合に有効か**

  AI システムがユーザからのアップロードファイルや Web サイト等の外部データを参照する場合、AI エージェントが連携先システムと接続する場合に有効である。外部データに仕込まれた不正指示により、AI システムが権限外の操作を実行してしまうリスクが高まる構成で重要度が高い。
- **なぜこの対策が有効か**
  - 対策 3 に対し、T-2 はシステム指示と入力データを構造的に分離することで攻撃の入口そのものを狭める効果がある。MCP サーバのツール定義改ざん（ツールポイズニング）のように、AI が信頼するツール定義に不正指示が混入するケースにも有効である。
- **残存リスク**
  - 未知の攻撃手法に対しては完全な防御を保証できない。
- **導入・運用上の留意点**
  - 現時点では多くの AI モデルがこの分離を十分に実現できておらず、対応可能な AI 提供者が限定される。本対策に過度に依存せず、対策 1～3 との多層防御の一部として位置づける必要がある。

- 調達仕様書に要件として盛り込み、AI 提供者の安全性テスト報告書の開示を求める。対応状況は AI 提供者の選定段階で確認する（第 3 章参照）。
- 外部データを参照する構成では、参照先データソースの信頼性を事前に評価し、不特定多数が書き込める外部ソースの参照は最小限にとどめることも有効な補完策となる。

(5) ネットワーク分離によるサンドボックス化

- **組み合わせる対策**
  - 技術要件：ユーザ環境分離とデータ処理環境の確保（T-5）
  - 運用対策：インフラ・環境分離（O-4）
- **どのような場合に有効か**

  AI エージェントが基幹システム等の重要なシステムと同一ネットワーク上にある場合、または AI がアクセスできるシステムの重要度が高い場合に有効である。
- **なぜこの組み合わせが有効か**

  T-5 により AI 提供者側での処理環境の隔離を求め、O-4 により自社ネットワーク内で AI システムを重要システムから切り離すことで、AI が侵害された場合の被害範囲を限定できる。
- **残存リスク**
  - 許可された通信経路内での不正操作は防げない。（対策 1 との併用が望ましい。）
- **導入・運用上の留意点**
  - ネットワーク構成の変更は情報システム部門との調整が必要であり、工数とコストが発生する。
  - AI システムから連携先システムへの接続が必要な場合でも、接続する API エンドポイントは最小限にし、その他の直接アクセスを遮断する設計が望ましい。

(6) 個別ルールによる利用範囲の制限

- **対象の対策**
  - 人的統制：個別システム特有のルール策定（H-2）
- **どのような場合に有効か**

  AI 利用者が連携先システムを自由に追加・変更できる AI システムの場合に有効である。
- **なぜこの対策が有効か**

  連携先システムの追加・変更をルールとして統制することで、セキュリティ部門が関与しないまま高リスクな連携が追加されることを防止できる。
- **残存リスク**

- 周知・教育が不十分な場合や、業務実態とルールが乖離している場合には、ルールが形骸化する恐れがある。
- 悪意のある内部不正によるルールの意図的な無視は防げない。

- **導入・運用上の留意点**
  - 新しい連携先システムを追加する場合は、セキュリティ部門の承認を必須とするプロセスを設ける。承認基準と手順を AI 管理者と連携して策定する。
  - ルール策定自体のコストは低く、専門スキルも不要であるが、承認プロセスの運用には担当者の工数が継続的に発生する。

■対策の進め方のまとめ

本発生被害への対策は、以下のステップで段階的に進めることを推奨する。

① 権限を最小限に設定し、人間の承認プロセスを整備する（対策 1～2）。

② 入出力のフィルタリングとログを有効化し、攻撃の遮断と追跡体制を整える（対策 3）。

③ 固有リスクの高さに応じて、プロンプトインジェクション対策、ネットワーク分離、個別ルールを追加する（対策 4～6）。

①②は AI エージェントを利用するすべての企業が実施すべき基本対策である。単体の生成 AI の場合は本被害のリスクは限定的であり、①②の対策で多くのケースにおいて十分である。なお対策 4（T-2）は AI 提供者の対応状況に依存するため実現が困難なケースも多く、対応可能な AI 提供者の選定基準として活用することが現実的である。

発生被害 C：システムや内部データの改ざん

改ざんは、AI エージェントがデータベースに対して書き込み権限を持つ場合にリスクが高まる。また、RAG の知識ベースへの悪意あるドキュメントの投入（データポイズニング）も改ざんの一形態である。改ざんされたデータが AI の回答に反映されると、後続の業務プロセスに被害が波及する点が特徴である。

■まず取り組むべき基本対策

（1） 書き込み権限の最小化

- **組み合わせる対策**
  - 技術要件：連携先システムにおける実行権限の最小化（T-6）
  - 運用対策：利用者のアクセス権限設定（O-2）
- **なぜこの組み合わせが有効か**

  T-6 で AI システムの書き込み権限を読み取り専用に制限し、O-2 で自社環境でも書き込み権限を付与しないことで、改ざん操作そのものを構造的に不可能にできる。
- **残存リスク**

- 読み取り権限のみでも、AI の出力を人が手動で転記することによる誤情報の後続業務への伝播は防げない。

- **導入・運用上の留意点**
  - 業務上、書き込みが必要な場合は対象テーブルや操作を限定し、次項の対策 2 を併用する。AI 利用者と書込可能範囲を協議する。
  - 既に社内で運用している ID 管理の仕組みを AI システムにも適用することで、実現可能な場合も多い。アクセス権限の設計は人事部門（異動情報）や各事業部門（業務上必要なアクセス範囲）との連携が不可欠である。

(2) 重要データ操作に対する人間の承認

- **対象の対策**
  - 人的統制：人間介入・承認プロセス（H-1）
- **なぜこの対策が有効か**

  業務上書き込み権限を付与せざるを得ない場合において、AI の書き込み操作実行前に人間の承認を求めることで不正な書き込み指示を遮断できる。

- **残存リスク**
  - 承認対象が多い場合、承認疲れにより形骸化する。
  - 操作の内容（SQL クエリ・API 呼び出し等）を承認者が技術的に理解できないケースがあり、実質的な確認なしに承認されるリスクがある。
- **導入・運用上の留意点**
  - 基幹データや RAG の知識ベースへの書き込みは承認必須とし、影響の小さいデータは AI の自律実行を許容することで、実効性を高めることができる。
  - 承認者がデータ操作の意味と影響範囲を理解できるよう、操作内容を平易に表示する工夫や説明資料の整備が望ましい。技術的な内容を含む操作については情報システム部門が承認に関与する体制も検討する。

(3) ログの有効化と入出力フィルタリング

- **組み合わせる対策**
  - 技術要件：入出力の監視とフィルタリング機能（T-4）
  - 運用対策：入出力のフィルタリング設定（O-1）
  - 技術要件：ログの取得（T-9）
  - 運用対策：ログの有効化（O-5）
- **なぜこの組み合わせが有効か**

T-4/O-1 により改ざんを目的とした不正な指示を入力段階で遮断し、T-9/O-5 によりフィルタリングの遮断履歴を含む変更前後の値・操作内容を記録することで、改ざんの予防・検知・原状回復を一体的に実現できる。

- **残存リスク**
  - ログ管理体制が不十分な場合は、ログの書き換え・消去による証拠隠滅のリスクがある。
  - フィルタリングで検知できない巧妙なプロンプトインジェクションは防げない。
- **導入・運用上の留意点**
  - まずはログの記録から始め SIEM 連携やリアルタイム分析は段階的に整備する方法でも良い。SIEM 連携やリアルタイム分析の実施には、ネットワークやセキュリティに関する専門知識が必要である。
  - ログの真正性を保護する必要がある。初期段階ではログの保存場所へのアクセス権限を最小化することから始めることが現実的である。

■固有リスクに応じて追加する対策

以下は、RAG の知識ベースを利用する場合や改ざんデータが連携先システムに自動連携される構成の場合に追加で検討する対策である。

（4） RAG の知識ベースに対する安全性検証

- **対象の対策**
  - 技術要件：参照データに対するアクセス制御と安全性検証（T-7）
- **どのような場合に有効か**

  RAG を活用して社内文書を AI に参照させる場合に有効である。特に信頼性の低い外部データを大量に取り込む場合や、AI を騙すための悪意ある記述がデータに混入するリスクがある場合に重要度が高い。
- **なぜこの対策が有効か**

  被害 A の対策 6 が RAG 知識ベースへのアクセス制御による情報漏洩防止を主眼とするのに対し、本対策は RAG 知識ベースへの悪意あるドキュメントの混入・改ざんの防止を主眼とする。AI 提供者に対して外部データに含まれる悪意ある命令や不正コードのサニタイズとデータのポイズニング検知を求めることで、知識ベース経由の改ざんリスクを低減できる。
- **残存リスク**
  - 巧妙に偽装されたポイズニングはサニタイズ処理でも検知できない場合がある。
- **導入・運用上の留意点**
  - AI が参照する社内データの追加・更新は、データの所管部門の承認を得た上で実施する。
  - AI 提供者のサニタイズ・ポイズニング検知機能の対応状況を調達段階で確認する。

- 外部データを RAG の知識ベースに取り込む場合は、信頼性の低いソースからの大量取り込みを避け、取り込み前のレビュープロセスを設けるべきである。レビューには業務知識のある担当者の関与が必要であり、継続的な工数が発生する。

(5) 個別ルールによるデータ操作の統制

- **対象の対策**
  - 人的統制：個別システム特有のルール策定（H-2）
- **どのような場合に有効か**

  改ざん後のデータが連携先システムや後続業務に自動連携・同期される構成の場合や、改ざんの影響が広く伝搬するリスクがある場合に有効である。
- **なぜこの対策が有効か**

  改ざんデータが連携先システムや後続業務に自動連携される構成では、技術的対策だけでは伝播を完全に防ぐことが難しい。本対策によって、改ざんデータが伝播する前に人間が介入できる機会を確保する。
- **残存リスク**
  - 周知・教育が不十分な場合や、業務実態とルールが乖離している場合には、ルールが形骸化する恐れがある。
  - 悪意のある内部不正によるルールの意図的な無視は防げない。
  - 巧妙に偽装されたデータの目視チェック漏れは排除できない。
- **導入・運用上の留意点**
  - ルールは AI システムの利用部門と連携して策定し、業務実態と乖離しないようにする。
  - ルールの実効性を技術的に補完するため、重要度の高いデータから段階的にデジタル署名やチェックサムによるデータ検証仕組みを導入することが望ましい。ただし実装には開発スキルが必要であり、情報システム部門または外部ベンダーとの連携が必要となる。
  - ルール策定自体のコストは低く、専門スキルも不要であるが、承認プロセスの運用には担当者の工数が継続的に発生する。

■対策の進め方のまとめ

本発生被害への対策は、以下のステップで段階的に進めることを推奨する。

① AI システムの書き込み権限を最小限にし、重要データの操作には人間の承認を必須とする（対策 1、2）。

② ログを有効化し、入出力フィルタリングを設定する（対策 3）。

③ RAG を利用する場合は、データ投入プロセスを整備し、改ざんの影響が波及する構成では個別ルールを策定する（対策 4、5）。

AI に書き込み権限を一切付与しない構成であれば、①のみで改ざんリスクの大部分を低減できる。

発生被害 D：AI システムのサービス停止

サービス停止は、AI 自体への攻撃だけでなく、AI が依存する外部サービスや AI 提供者の障害によっても発生する。AI を補助的な用途に利用しており、停止した場合も業務が進行可能であれば、基本対策で十分なケースが多い。

■まず取り組むべき基本対策

（1） リソース制限の設定

- **組み合わせる対策**
  - 技術要件：リソース消費の制御（T-8）
  - 運用対策：リソース・コスト管理（O-3）
- **なぜこの組み合わせが有効か**

  T-8 で AI 提供者にリクエスト数・トークン数・処理時間等に対するハードリミット機能と異常検知による自動停止機能を求め、O-3 で自社の利用状況に応じた閾値を設定することで、大量リクエストや AI エージェントの処理暴走による無限ループによるリソース枯渇起因のサービス停止を防止できる。
- **残存リスク**
  - 閾値が低すぎることによる正常業務へ影響が出る。
  - 複数拠点からの分散型攻撃（小口リクエストの乱発）に対しては、回数上限を意図的に回避される可能性がある。
  - AI 提供者側の障害や計画外メンテナンス、外部サービスの停止によるサービス停止には効果がない。
- **導入・運用上の留意点**
  - 閾値は実際の利用状況を見ながら段階的に最適化する。初期段階では余裕を持った閾値設定から始めることが望ましい。
  - 多くの AI システムでは管理画面から設定可能であり、追加の開発作業は不要な場合が多い。

（2） ログの有効化による AI システムの監視

- **組み合わせる対策**
  - 技術要件：ログの取得（T-9）
  - 運用対策：ログの有効化（O-5）
- **なぜこの組み合わせが有効か**

  AI システムの稼働状況・応答時間の遅延・エラーの発生状況等を記録・検知することで、停止の兆候の早期発見と停止発生後の原因分析や再発防止に活用できる。

- **残存リスク**
  - ログ分析体制が不十分な場合は兆候検知が遅延し、停止を未然に防げない。
  - AI 提供者側の障害によるサービス停止は本対策では検知できない。
- **導入・運用上の留意点**
  - まずはログを記録することから始め、リアルタイム検知は段階的に整備する。
  - AI 開発者および AI 提供者の SLA 確認と信頼性評価については第 3 章のライフサイクル・調達フェーズで対応する。

■固有リスクに応じて追加する対策

以下は、AI システム停止が業務プロセスやサービス遅延・破綻を招く場合に追加で検討する対策である。

(3) ネットワーク分離と環境保護

- **対象の対策**
  - 運用対策：インフラ・環境分離（O-4）
- **どのような場合に有効か**

  外部からの DoS 攻撃等により AI システムが停止するリスクが高い構成の場合や、本番環境と検証環境が分離できない場合に有効である。
- **なぜこの対策が有効か**

  本対策により、AI システムへの外部からの直接アクセスをプロキシ経由に限定し、本番環境と検証環境を分離することで、DoS 攻撃等の外部攻撃によるサービス停止リスクを低減できる。また本番環境と検証環境を分離することで、検証作業が本番環境の稼働に影響するリスクも排除できる。
- **残存リスク**
  - 本番環境の接続ポイントが漏洩した場合に、直接的な DoS 攻撃によるサービス停止が生じる可能性がある。
  - ネットワーク分離はサービス停止を防ぐ対策であり、停止が発生した場合の業務継続には別途対応が必要である。AI 業務依存度が高い場合は、手動代替手順やフォールバック手順をあわせて整備することが望ましい。
- **導入・運用上の留意点**
  - ネットワーク構成の変更は情報システム部門との調整が必要であり、工数とコストが発生する。

■対策の進め方のまとめ

本発生被害への対策は、以下のステップで段階的に進めることを推奨する。

① リソース制限を設定し、ログ取得を有効化する（対策 1、2）。

② ネットワーク分離や脆弱性管理の体制を整備する（対策 3）。

AI を補助的な用途に利用しており停止しても業務が進行可能な場合は、①で十分なケースが多い。AI システムの業務依存度が高い場合は、AI システムが停止した際の手動代替手順やフォールバック手順をあわせて整備することが望ましい。

発生被害 E：虚偽や低品質な出力の利用および出力の誤用

現時点ではいかなる AI モデルもハルシネーションを完全に排除することができない。AI の出力には誤りが含まれる可能性があることを前提とした対策が必要であり、人間の確認プロセスが対策の中核となる。

■まず取り組むべき基本対策

（1） 人間の確認プロセスの整備

- **組み合わせる対策**
  - 人的統制：人間介入・承認プロセス（H-1）
  - 人的統制：個別システム特有のルール策定（H-2）
- **なぜこの組み合わせが有効か**

  H-1 により「AI の出力は下書きであり、最終判断は人間が行う」という原則を業務プロセスで制度化し、H-2 により当該システムにおける確認の具体的な手順等を定める。他の発生被害では H-1 と H-2 を別々の対策として扱っているが、本対策では AI の出力確認という一つの業務行為に対して承認フロー（H-1）と確認手順（H-2）が一体で機能するため、組み合わせて一つの対策として位置づける。両者を組み合わせることで、ハルシネーションの見落としと AI 出力の無批判な誤用の両方を防止できる。
- **残存リスク**
  - AI 利用者の専門知識が不足している場合、AI の誤りを見抜けない。
  - 大量出力への対応や業務繁忙期に確認が形骸化する恐れがある。
- **導入・運用上の留意点**
  - 確認対象を業務リスクに応じて絞り込み、すべての出力に同じレベルの確認を求めないことで、実効性が高まる。
  - AI 利用者に対してハルシネーションの特性と確認方法に関する教育・トレーニングを実施することで、専門知識不足による見落としリスクを低減できる。教育については第 2 章を参照されたい。
  - ルール策定自体のコストは低く、追加の開発作業も不要であり、すべての企業が実施可能な対策である。
  - 重要システムにおいては、出力結果を毎日確認し、ハルシネーションが発生していないか等を確認すべきである。

(2) 出力のフィルタリング

- **組み合わせる対策**
  - 技術要件：入出力の監視とフィルタリング機能（T-4）
  - 運用対策：入出力のフィルタリング設定（O-1）
- **なぜこの組み合わせが有効か**

  AI 提供者の汎用フィルターに自社のポリシーを追加設定することで、差別的表現や権利侵害等の明らかに不適切な出力を遮断できる。
- **残存リスク**
  - 巧妙に偽装された不適切な出力はフィルタリングをすり抜ける可能性がある。
  - ハルシネーションはフィルタリングでは検知できない。（対策 1 との併用が望ましい。）
  - フィルタリングが厳しすぎる場合、有用な回答まで遮断される過検知の恐れがある。
- **導入・運用上の留意点**
  - 多くの AI システムでは管理画面からコンテンツフィルターの設定が可能であり、追加の開発作業は不要な場合が多い。
  - 過検知と検知漏れのバランスを調整しながら段階的にフィルタリングルールを整備する。自社のコンテンツポリシーに合わせたカスタムルールの設定にはセキュリティ部門との連携が必要である。

■固有リスクに応じて追加する対策

以下は、AI の出力が重要な業務判断に利用される場合や顧客・取引先・社会全体に影響を及ぼし得る場合に追加で検討する対策である。

(3) 推論プロセスの可視化

- **組み合わせる対策**
  - 技術要件：ログの取得（T-9）
  - 運用対策：ログの有効化（O-5）
- **どのような場合に有効か**

  AI の出力を重要な業務判断に利用する場合（融資判定、法的見解、顧客提案等）や、回答の根拠確認が求められる場合に有効である。
- **なぜこの対策が有効か**

  本対策より取得するログには AI の推論プロセスが含まれる。このログを活用して推論プロセスを可視化することで、AI の回答がどのような根拠に基づいているかを確認でき、ハルシネーションや誤った推論を発見しやすくなる。
- **残存リスク**

- 推論プロセスのログに記録された説明自体がハルシネーションを含む可能性があること、またログが AI の内部計算過程を完全に反映しているとは限らないことから、可視化による確認には本質的な限界がある。

- **導入・運用上の留意点**
  - AI 提供者の可視化機能の対応状況を調達時に確認する。
  - 推論プロセスの詳細なログ取得はトークン消費・ストレージコストが増大するため、すべての出力に適用するのではなく、重要な業務判断に利用する出力に限定して取得することが現実的である
  - 推論プロセスの解釈には AI の動作原理に関する一定の知識が必要であり、担当者への教育が必要となる場合がある。

（4） 出力品質の定期評価

- **対象の対策**
  - 技術要件：真正性・透明性の確保（T-3）
- **どのような場合に有効か**

  AI を継続利用しており、品質劣化のリスクがある場合や、AI 生成物を外部に公開・提出する場合に有効である。
- **なぜこの対策が有効か**

  AI 提供者にモデルの構成情報・更新履歴の開示を求めることで、品質変動の原因を追跡できる。あわせて定期的なレッドチーミングやベンチマークテスト・サンプリング評価を実施することで、品質劣化を早期に検知できる。なお被害 A の対策 8 がサプライチェーンリスクの可視化を主眼とするのに対し、本対策はモデルの透明性を出力品質の継続的な監視に活用する点で目的が異なる。
- **残存リスク**
  - 評価用データセットがカバーしない領域の品質劣化は検知できない。
- **導入・運用上の留意点**
  - AI を敵に見立ててわざと悪意ある指示を入力し、安全ガードレールを突破できるかを検証する手法（レッドチーミング）の実施には AI の攻撃手法の知識が必要であり、外部専門家への委託も選択肢に入る。
  - 評価用データセットの整備・ベンチマークテストの設計には業務知識と AI 評価の専門知識が必要であり、セキュリティ部門と AI 利用部門の連携が必要となる。

■対策の進め方のまとめ

本発生被害への対策は、以下のステップで段階的に進めることを推奨する。

① 人間の確認プロセスを整備し、個別ルールを策定する（対策 1）。コストをかけずに実施できる最も効果的な対策である。

② AI 提供者のフィルタリング機能を有効化する（対策 2）。

③ 利用用途の重要度に応じて、推論プロセスの可視化、品質の定期評価を追加する（対策 3、4）。

本被害は AI の技術的制約に起因するものであり、技術的対策だけで完全に解決することは困難である。「AI に任せきりにしない」という組織文化の醸成が最も重要である。

発生被害 F：第三者への加害

本被害は、自社の AI システムが攻撃の加害者になるという点で、他の発生被害とは性質が異なる。AI システムが悪用されにくくする対策と、悪用された場合の加害の及ぶ範囲を限定する対策を組み合わせることが重要である。特に AI システムが社内および社外のシステムに接続する環境でリスクが高くなる。

■まず取り組むべき基本対策

（1） 権限の最小化と入出力フィルタリング

- **組み合わせる対策**
  - 技術要件：連携先システムにおける実行権限の最小化（T-6）
  - 技術要件：入出力の監視とフィルタリング機能（T-4）
  - 運用対策：入出力のフィルタリング設定（O-1）
  - 運用対策：利用者のアクセス権限設定（O-2）
- **なぜこの組み合わせが有効か**

  T-6 で、AI システム自身が連携先システムに対して持つ実行権限を必要最小限に抑える。例えば情報検索のみを目的とする AI システムには読取専用権限のみを付与し、書き込み、削除、外部送信、権限変更等の操作を許可しない構成とすることで、AI システムが掌握された場合でも、重大な操作が実行されにくい状態にできる。

  T-4/O-1 の入出力フィルタリングにより、AI を掌握するための攻撃（プロンプトインジェクション等）や、第三者への加害につながる指示を入力段階で検知・遮断する。フィルタリングルールには禁止語だけでなく、外部システムへの攻撃、過剰リクエスト送信、認証情報の取得、データ削除、権限変更、外部送信等を誘導する指示を含めることが望ましい。

  O-2 で利用者ごとに AI システムを通じて実行できる操作範囲を制御する。これにより、AI エージェント側の権限と利用者側の権限の双方から、加害につながる操作を制限できる。
- **残存リスク**
  - 最小権限の範囲内で実行可能な操作を悪用した加害（正規に許可された連携先システムへの過剰リクエスト送信等）は防げない。

- フィルタリングルールに未登録の新たな攻撃手法や、巧妙に偽装された指示によるプロンプトインジェクションは検知漏れが生じる。
- 正規利用者の権限を悪用する攻撃や、連携先システム側の脆弱性を突く攻撃については、AI システム側のフィルタリングだけでは十分に防げない。

- **導入・運用上の留意点**
  - 連携先システムごとの実行権限は、AI 導入者および AI 管理者が情報システム部門と連携して行い、連携先システムへの影響範囲を事前に把握する。また、定期的に見直しを実施する。
  - フィルタリングルールの継続的な更新が必要である。

（2） ログの有効化と人間の承認

- **組み合わせる対策**
  - 技術要件：ログの取得（T-9）
  - 運用対策：ログの有効化（O-5）
  - 人的統制：人間介入・承認プロセス（H-1）
- **なぜこの組み合わせが有効か**

  T-9/O-5 によってログから連携先システムへの異常なアクセスパターンを検知し、H-1 の人間の承認により重要な外部操作の不正実行を阻止できる。ログによる検知だけでは攻撃実行後の対応になるが、H-1 を組み合わせることで実行前に阻止できる点が本対策の核心であり、早期検知と実行前阻止の二段構えで加害行為を防止する。
- **残存リスク**
  - ログ分析体制が不十分な場合、加害行為の検知が遅延、または被害が拡大する恐れがある。
  - 承認対象の操作が多すぎると承認者の負荷が高まり、内容を十分に確認しないまま承認する「承認疲れ」により、承認プロセスが形骸化する恐れがある。
- **導入・運用上の留意点**
  - 監視ルールの設計はセキュリティ部門と連携する。
  - 異常を検知した場合の対応フローは AI 管理者および情報システム部門と連携する。
  - すべての操作に承認を求めると業務効率と承認精度の双方が低下するため、連携先システムへの重要操作に限定して承認を求める。

■固有リスクに応じて追加する対策

以下対策 3 は、AI システムが重要な連携先システムと接続する場合やサプライチェーン経由の攻撃リスクが高い場合に追加で検討する対策である。

（3） プロンプトインジェクション対策の強化とネットワーク分離

- **組み合わせる対策**
  - 技術要件：指示とデータの構造的分離（T-2）
  - 運用対策：インフラ・環境分離（O-4）
- **どのような場合に有効か**

  AI システムが外部データを参照して動作する場合や AI システムと連携先システムが同一ネットワーク上にある場合に有効である。
- **なぜこの組み合わせが有効か**

  O-4 のネットワーク分離により、万一 AI が掌握された場合でも、連携先以外の重要システムへの横展開を遮断し、被害範囲を限定する。なお、T-2 の指示とデータの構造的分離は、外部データに紛れ込んだ不正な指示を AI が実行してしまうこと（プロンプトインジェクションによる掌握）を困難にする技術である。ただし現時点では多くの AI モデルがこの分離を十分に実現できておらず AI システムに広く組み込まれるに至っていないため、AI 提供者が当該機能を提供している場合には積極的に活用することが望ましい。
- **残存リスク**
  - 許可された通信経路内での攻撃は、ネットワーク分離を行っても遮断できない。
- **導入・運用上の留意点**
  - ネットワーク構成の変更は情報システム部門との調整が必要。AI 提供者の攻撃耐性は選定段階で確認する（第 3 章参照）。

■対策の進め方のまとめ

本発生被害への対策は、以下のステップで段階的に進めることを推奨する。

① 権限の最小化とフィルタリングにより悪用の被害範囲を限定し、ログと承認プロセスで検知・阻止する（対策 1、2）。

② 第 5 章で評価した固有リスクの高さに応じて、プロンプトインジェクション対策とネットワーク分離により、AI の掌握と横展開を防止する（対策 3）。

本被害は自社が加害者になるリスクであり、社会的信用の毀損に直結する。AI システムを連携先システムと接続させる場合は、①の対策を確実に実施することが重要である。

発生被害 G：過剰課金の発生

過剰課金は、AI エージェントの処理暴走や EDoS 攻撃、認証情報の漏洩等により発生する。

■まず取り組むべき基本対策

（1） コスト監視とアラートの設定

- **組み合わせる対策**

- 運用対策：リソース・コスト管理（O-3）
- 技術要件：ログの取得（T-9）
- 運用対策：ログの有効化（O-5）

- **なぜこの組み合わせが有効か**

  O-3 により API 等の利用額に閾値を設定し、超過時にアラート発報を行う仕組みを整備し、T-9/O-5 により利用状況を記録することで、コスト暴走の早期検知や攻撃の事後追跡が実現できる。

- **残存リスク**
  - 検知から停止までのタイムラグ中にコストが発生する。
  - 閾値設定が不適切な場合、正常利用時にもアラートが頻発し形骸化する恐れがある。
  - 閾値を下回る小口の不正利用は検知が困難である。

- **導入・運用上の留意点**
  - 多くの AI システムでは管理画面からこれらの設定が可能であり、追加の開発作業は不要な場合が多い。
  - 利用額に段階的な閾値（例：警告レベル、停止レベル）を設定する。閾値は実際の利用状況を見ながら調整する。
  - アラート発報時の対応フローを事前に整備する必要がある。

(2) アクセス制御による不正利用の防止

- **対象の対策**
  - 運用対策：利用者のアクセス権限設定（O-2）

- **なぜこの対策が有効か**

  多要素認証（MFA）の導入によりアカウントの不正アクセスを防止し、API キーのシークレット管理ツールへの格納・定期的なキーローテーションにより、認証情報の漏洩リスクを低減することで、不正利用起因の過剰課金を防止できる。

- **残存リスク**
  - MFA 疲労攻撃（大量の MFA 認証要求を送りつけて誤承認を誘発する攻撃）により認証を突破される恐れがある。
  - キーローテーションが未実施または間隔が長い場合、漏洩した認証情報が長期間悪用されるリスクがある

- **導入・運用上の留意点**
  - キーローテーションの自動化基盤の構築が望ましいが、追加の開発工数を要する。自動化が困難な場合は定期的な手動ローテーションのスケジュールを設定するとよい。

■固有リスクに応じて追加する対策

以下は、AI システムの利用規模が大きくコスト暴走時の影響が甚大な場合に追加で検討する対策である。

(3) リソース消費のハードリミット設定

- **組み合わせる対策**
  - 技術要件：リソース消費の制御（T-8）
  - 運用対策：リソース・コスト管理（O-3）
- **どのような場合に有効か**

  AI システムの利用規模が大きくコスト暴走時の影響が甚大な場合や、AI エージェントを利用しており処理暴走のリスクが高い場合に有効である。
- **なぜこの組み合わせが有効か**

  対策 1 のコスト監視が「閾値を超えたら止める」のに対し、本対策は T-8 で AI 提供者に対して、リクエスト数・トークン数、処理時間等に対するハードリミット機能の実装を求めることで、そもそも閾値を超えられないようにする。
- **残存リスク**
  - 閾値設定が低すぎると正常な業務に影響が出る。
- **導入・運用上の留意点**
  - 1 リクエスト当たりの最大トークン数（max tokens）や一定時間あたりの API リクエスト数（RPM）およびトークン処理量（TPM）等を使用用途に応じて設定することが望ましい。閾値は実際の利用状況を見ながら段階的に調整する。

■対策の進め方のまとめ

本発生被害への対策は、以下のステップで段階的に進めることを推奨する。

① コスト監視とアラートを設定し、アクセス制御で不正利用を防止する。（対策 1、2）。多くの AI システムでは管理画面から設定可能であり、AI システムの利用を開始する際に必ず実施する。

② AI エージェントを利用する場合や利用規模が大きい場合は、AI 提供者にリソース消費のハードリミット機能を求める。（対策 3）。

単体の生成 AI の場合は、①の基本対策で大部分のリスクを低減できる。