



# Artificial Intelligence Security Verification Standard

Initial Version Work In Progress

, 2026

# Table of Contents

<b>扉絵</b> .....	1
標準について .....	1
著作権およびライセンス .....	1
プロジェクトリーダー .....	2
寄稿者とレビュー .....	2
<b>序文</b> .....	3
はじめに .....	3
AISVS バージョン 1.0 の主要目標 .....	3
明確に定義された範囲 .....	3
<b>AISVSを使用して</b> .....	4
人工知能セキュリティ検証レベル .....	4
レベルの定義 .....	4
役割 (D/V) .....	4
<b>C1 トレーニングデータガバナンス &amp; バイアス管理</b> .....	6
制御目標 .....	6
C1.1 トレーニングデータの出所 .....	6
C1.2 トレーニングデータのセキュリティと整合性 .....	6
C1.3 トレーニングデータのラベリング品質、完全性、およびセキュリティ .....	7
C1.4 トレーニングデータの品質とセキュリティ保証 .....	7
C1.5 データ系譜とトレーサビリティ .....	8
参考文献 .....	8
<b>C2 ユーザー入力検証</b> .....	9
制御目標 .....	9
C2.1 プロンプトインジェクション防御 .....	9
C2.2 敵対的サンプルへの耐性 .....	9
C2.3 プロンプト文字セット .....	10
C2.4 スキーマ、型および長さの検証 .....	10
C2.5 コンテンツ&ポリシースクリーニング .....	11
C2.6 入力レート制限と乱用防止 .....	11
C2.7 マルチモーダル入力検証 .....	12
C2.8 リアルタイム適応型脅威検出 .....	12
参考文献 .....	13
<b>C3 モデルライフサイクル管理と変更管理</b> .....	14
制御目標 .....	14
C3.1 モデル認証と整合性 .....	14
C3.2 モデル検証およびテスト .....	14
C3.3 制御された展開とロールバック .....	15

C3.4 セキュア開発プラクティス . . . . .	15
C3.5 モデルの引退と廃止 . . . . .	16
参考文献 . . . . .	16
<b>C4 インフラストラクチャ、構成および展開のセキュリティ . . . . .</b>	<b>17</b>
制御目標 . . . . .	17
C4.1 ランタイム環境の分離 . . . . .	17
C4.2 セキュアなビルド&デプロイメントパイプライン . . . . .	17
C4.3 ネットワークセキュリティとアクセス制御 . . . . .	18
C4.4 シークレットおよび暗号鍵管理 . . . . .	18
C4.5 AI ワークロードのサンドボックス化と検証 . . . . .	19
C4.6 AI インフラストラクチャリソース管理、バックアップおよびリカバリー . . . . .	19
C4.7 AI ハードウェアセキュリティ . . . . .	20
C4.8 エッジおよび分散AIセキュリティ . . . . .	20
参考文献 . . . . .	21
<b>AIコンポーネントおよびユーザーのためのC5アクセス制御とアイデンティティ管理 . . . . .</b>	<b>22</b>
制御目標 . . . . .	22
C5.1 アイデンティティ管理と認証 . . . . .	22
C5.2 認可とポリシー . . . . .	22
C5.3 クエリ時のセキュリティ強制 . . . . .	23
C5.4 出力フィルタリングとデータ損失防止 . . . . .	23
C5.5 マルチテナント分離 . . . . .	24
C5.6 自律エージェントの認可 . . . . .	24
参考文献 . . . . .	24
<b>C6 モデル、フレームワーク、およびデータのサプライチェーンセキュリティ . . . . .</b>	<b>26</b>
制御目標 . . . . .	26
C6.1 事前学習済みモデルの審査および起源の完全性 . . . . .	26
C6.2 フレームワーク&ライブラリスキヤンニング . . . . .	26
C6.3 依存関係の固定と検証 . . . . .	27
C6.4 信頼できるソースの強制措置 . . . . .	27
C6.5 サードパーティデータセットリスク評価 . . . . .	28
C6.6 サプライチェーン攻撃モニタリング . . . . .	28
C6.7 モデルアーキテクチャクトのための ML-BOM . . . . .	29
参考文献 . . . . .	29
<b>C7 モデルの動作、出力制御および安全保証 . . . . .</b>	<b>30</b>
制御目標 . . . . .	30
C7.1 出力フォーマットの強制 . . . . .	30
C7.2 幻覚検出と緩和 . . . . .	30
C7.3 出力の安全性およびプライバシーフィルタリング . . . . .	31
C7.4 出力およびアクション制限 . . . . .	31
C7.5 出力の説明可能性 . . . . .	32

C7.6 監視統合 . . . . .	32
7.7 生成メディアの安全対策 . . . . .	32
参考文献 . . . . .	33
<b>C8 メモリ、埋め込み、ベクターデータベースのセキュリティ . . . . .</b>	<b>34</b>
制御目標 . . . . .	34
C8.1 メモリおよびRAGインデックスへのアクセス制御 . . . . .	34
C8.2 埋め込みのサニタイズおよび検証 . . . . .	34
C8.3 メモリの有効期限、取り消し、および削除 . . . . .	35
C8.4 埋め込み反転および漏洩の防止 . . . . .	35
C8.5 ユーザー固有メモリのスコープ強制 . . . . .	36
C8.6 高度なメモリシステムセキュリティ . . . . .	36
参考文献 . . . . .	37
<b>9 自律的オーケストレーションとエージェント行動のセキュリティ . . . . .</b>	<b>38</b>
制御目標 . . . . .	38
9.1 エージェントのタスク計画と再帰予算 . . . . .	38
9.2 ツールプラグインのサンドボックス化 . . . . .	38
9.3 自律ループとコスト制限 . . . . .	39
9.4 プロトコルレベルの誤用防止 . . . . .	39
9.5 エージェントの識別と改ざん防止 . . . . .	40
9.6 マルチエージェントウォームリスク削減 . . . . .	40
9.7 ユーザーおよびツールの認証／認可 . . . . .	41
9.8 エージェント間通信のセキュリティ . . . . .	41
9.9 意図検証と制約強制 . . . . .	41
9.10 エージェント推論戦略のセキュリティ . . . . .	42
9.11 エージェントライフサイクル状態管理とセキュリティ . . . . .	43
9.12 ツール統合セキュリティフレームワーク . . . . .	43
C9.13 モデルコンテキストプロトコル(MCP)セキュリティ . . . . .	44
コンポーネントの整合性とサプライチェーンの衛生管理 . . . . .	44
認証と認可 . . . . .	44
安全なトランスポートとネットワーク境界防御 . . . . .	44
スキーマ、メッセージ、および入力検証 . . . . .	45
アウトバウンドアクセスとエージェント実行の安全性 . . . . .	45
輸送制限と高リスク境界管理 . . . . .	45
参考文献 . . . . .	46
<b>10 敵対的ロバストネスとプライバシー防御 . . . . .</b>	<b>47</b>
制御目標 . . . . .	47
10.1 モデルの整合性と安全性 . . . . .	47
10.2 敵対的事例の耐性強化 . . . . .	47
10.3 メンバーシップ推論緩和 . . . . .	48
10.4 モデル反転耐性 . . . . .	48

10.5 モデル抽出防御 . . . . .	48
10.6 推論時の毒されたデータ検出 . . . . .	49
10.7 動的セキュリティポリシー適応 . . . . .	49
10.8 リフレクションベースのセキュリティ分析 . . . . .	50
10.9 進化と自己改善のセキュリティ . . . . .	50
参考文献 . . . . .	50
<b>11 プライバシー保護と個人データ管理 . . . . .</b>	<b>52</b>
制御目標 . . . . .	52
11.1 匿名化とデータ最小化 . . . . .	52
11.2 忘れられる権利と削除の強制 . . . . .	52
11.3 差分プライバシーの保護措置 . . . . .	52
11.4 目的制限およびスコープクリープ保護 . . . . .	53
11.5 同意管理と合法的根拠のトラッキング . . . . .	53
11.6 プライバシー制御を伴う連合学習 . . . . .	53
参考文献 . . . . .	54
<b>C12 監視、ログ記録、および異常検知 . . . . .</b>	<b>55</b>
制御目標 . . . . .	55
C12.1 要求および応答のログ記録 . . . . .	55
C12.2 悪用検出および警告 . . . . .	55
C12.3 モデルドリフト検出 . . . . .	56
C12.4 パフォーマンスおよび動作のテレメトリ . . . . .	56
C12.5 AI インシデント対応計画および実行 . . . . .	56
C12.6 AIパフォーマンス劣化検出 . . . . .	57
C12.7 DAG 可視化とワークフローセキュリティ . . . . .	57
C12.8 プロアクティブなセキュリティ行動モニタリング . . . . .	57
参考文献 . . . . .	58
<b>C13 人間の監督、説明責任、およびガバナンス . . . . .</b>	<b>59</b>
制御目標 . . . . .	59
C13.1 キルスイッチおよびオーバーライドメカニズム . . . . .	59
C13.2 ヒューマン・イン・ザ・ループ意思決定チェックポイント . . . . .	59
C13.3 責任の連鎖と監査可能性 . . . . .	60
C13.4 説明可能なAI技術 . . . . .	60
C13.5 モデルカードと使用開示 . . . . .	60
C13.6 不確実性の定量化 . . . . .	61
C13.7 ユーザ向け透明性レポート . . . . .	61
参考文献 . . . . .	61
<b>付録A：用語集 . . . . .</b>	<b>62</b>
<b>付録B: 参考文献 . . . . .</b>	<b>67</b>
未完了事項 . . . . .	67
<b>付録C : AIセキュリティガバナンスと文書化（再編成） . . . . .</b>	<b>68</b>

目的 . . . . .	68
AC.1 AIリスク管理フレームワークの採用 . . . . .	68
AC.2 AI セキュリティポリシーと手順 . . . . .	68
AC.3 AIセキュリティの役割と責任 . . . . .	68
AC.4 倫理的なAIガイドラインの施行 . . . . .	69
AC.5 AI 規制遵守モニタリング . . . . .	69
AC.6 トレーニングデータのガバナンス、ドキュメント化とプロセス . . . . .	69
AC.6.1 データソーシングとデューデリジェンス . . . . .	69
AC.6.2 バイアスと公平性の管理 . . . . .	69
AC.6.3 ラベリングおよび注釈のガバナンス . . . . .	70
AC.6.4 データセット品質ゲートおよび検疫 . . . . .	71
AC.6.5 脅威/毒性検出およびドリフト . . . . .	71
AC.6.6 削除、同意、権利、保持およびコンプライアンス . . . . .	71
AC.6.7 バージョニングおよび変更管理 . . . . .	72
AC.6.8 合成データガバナンス . . . . .	72
AC.6.9 アクセスモニタリング . . . . .	72
AC.6.10 敵対的トレーニングガバナンス . . . . .	72
AC.7 モデルライフサイクルガバナンスとドキュメント管理 . . . . .	73
AC.8 プロンプト、入力、および出力の安全性ガバナンス . . . . .	73
AC.8.1 プロンプトインジェクション防御 . . . . .	73
AC.8.2 敵対的サンプル耐性 . . . . .	73
AC.8.3 コンテンツおよびポリースクリーニング . . . . .	74
AC.8.4 入力レート制限と悪用防止 . . . . .	74
AC.8.5 入力の出所と帰属 . . . . .	74
AC.9 マルチモーダル検証、MLOpsおよびインフラストラクチャガバナンス . . . . .	74
AC.9.1 マルチモーダルセキュリティ検証パイプライン . . . . .	74
AC.9.2 CI/CDおよびビルドセキュリティ . . . . .	74
AC.9.3 コンテナおよびイメージのセキュリティ . . . . .	74
AC.9.4 監視、警報、およびSIEM . . . . .	75
AC.9.5 脆弱性管理 . . . . .	75
AC.9.6 設定およびドリフト制御 . . . . .	75
AC.9.7 本番環境の強化 . . . . .	75
AC.9.8 リリースプロモーションゲート . . . . .	75
AC.9.9 ワークロード、容量、およびコスト監視 . . . . .	75
AC.9.10 承認と監査証跡 . . . . .	76
AC.9.11 IaC ガバナンス . . . . .	76
AC.9.12 非本番環境におけるデータ処理 . . . . .	76
AC.9.13 バックアップと災害復旧 . . . . .	76
AC.9.14 コンプライアンス&ドキュメンテーション . . . . .	76
AC.9.15 ハードウェアおよびサプライチェーン . . . . .	77

AC.9.16 クラウド戦略とポータビリティ . . . . .	77
AC.9.17 GitOps と自己修復 . . . . .	77
AC.9.18 ゼロトラスト、エージェント、プロビジョニングおよびレジデンシー証明 . . . . .	77
AC.9.19 アクセス制御とアイデンティティ . . . . .	77
上に統合される新しいアイテム . . . . .	78
<b>付録D: AI支援によるセキュアコーディングのガバナンスと検証 . . . . .</b>	<b>79</b>
目的 . . . . .	79
AD.1 AI支援セキュアコーディングワークフロー . . . . .	79
AD.2 AIツール資格認定と脅威モデリング . . . . .	79
AD.3 セキュアなプロンプトおよびコンテキスト管理 . . . . .	79
AD.4 AI生成コードの検証 . . . . .	80
AD.5 コード提案の説明可能性と追跡可能性 . . . . .	80
AD.6 繙続的フィードバックとモデル微調整 . . . . .	81
参考文献 . . . . .	81
<b>付録E: ツールとフレームワークの例 . . . . .</b>	<b>82</b>
目的 . . . . .	82
AE.1 トレーニングデータガバナンスとバイアスマネジメント . . . . .	82
AE.2 ユーザー入力の検証 . . . . .	82
<b>付録 B: 戰略的コントロール . . . . .</b>	<b>83</b>
C4.15 量子耐性インフラストラクチャセキュリティ . . . . .	83
C4.17 ゼロ知識インフラストラクチャ . . . . .	83
C4.18 サイドチャネル攻撃防止 . . . . .	84
C4.19 ニューロモルフィックおよび特殊AIハードウェアのセキュリティ . . . . .	84
C4.20 プライバシー保護計算基盤 . . . . .	85

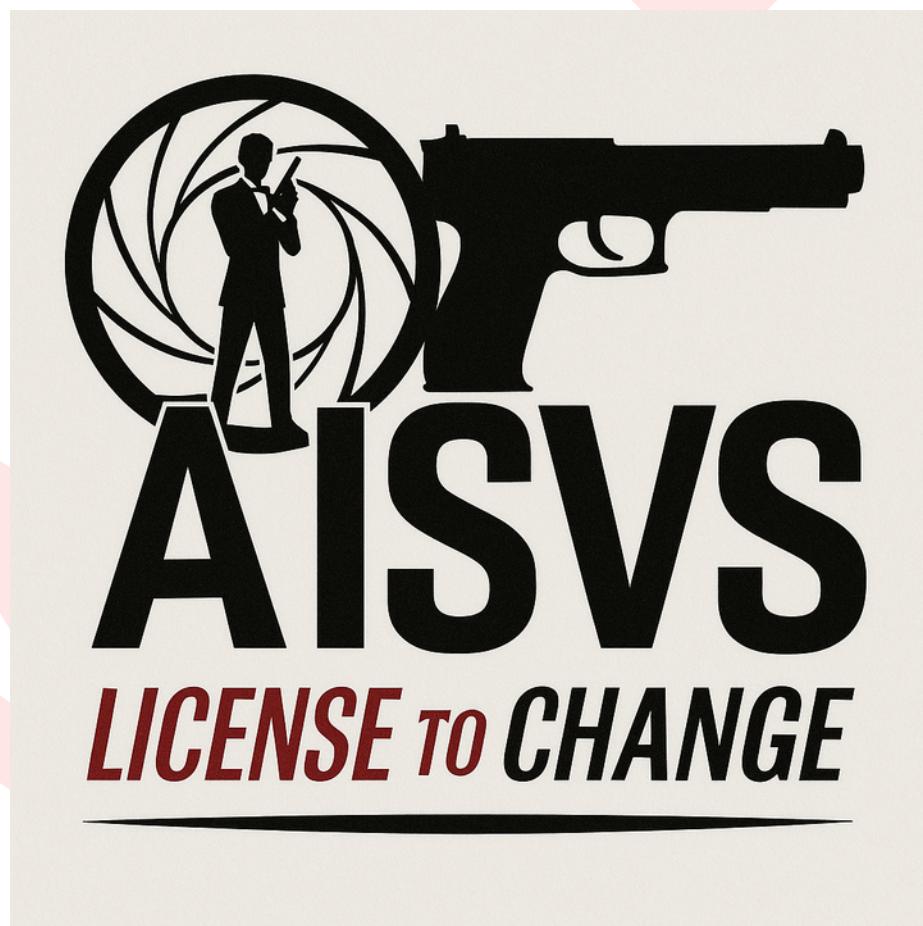
# 扉絵

## 標準について

人工知能セキュリティ検証標準(AISVS)は、データサイエンティスト、MLOpsエンジニア、ソフトウェアアーキテクト、開発者、テスター、セキュリティ専門家、ツールベンダー、規制当局、および消費者が、信頼できるAI対応システムおよびアプリケーションを設計、構築、テスト、および検証するために利用できるコミュニティ主導のセキュリティ要件カタログです。これは、データ収集やモデル開発から展開および継続的な監視に至るAIライフサイクル全体にわたるセキュリティコントロールを指定するための共通言語を提供し、組織がAIソリューションの回復力、プライバシー、および安全性を測定し向上させることを可能にします。

## 著作権およびライセンス

バージョン0.1(最初の公開ドラフト - 作業進行中)、2025



著作権 © 2025 The AISVS Project.

以下の条件のもとリリースされています Creative Commons Attribution-ShareAlike 4.0

## International License.

この作品を再利用または配布する場合、必ずライセンス条件を他者に明確に伝える必要があります。

## プロジェクトリーダー

ジム・マニコ

アラス “ラツス” メミシャジチ

## 寄稿者とレビュー者

<https://github.com/ottosulin>

<https://github.com/mbhatt1>

<https://github.com/vineethsai>

<https://github.com/cciprofm>

<https://github.com/deepakrpandey12>

AISVSは、人工知能システムの独自のセキュリティ課題に対応するために特別に作られた全く新しい標準です。より広範なセキュリティのベストプラクティスから着想を得ていますが、AISVSのすべての要件はAIの脅威の状況を反映し、組織がより安全で堅牢なAIソリューションを構築できるように、ゼロから開発されています。

# 序文

人工知能セキュリティ検証標準(AISVS)バージョン1.0へようこそ!

## はじめに

2025年に共同コミュニティの取り組みにより設立されたAISVSは、最新のAIモデル、パイプライン、およびAI対応サービスの設計、開発、展開、および運用時に考慮すべきセキュリティ要件を定義しています。

AISVS v1.0は、プロジェクトリーダー、作業グループ、および広範なコミュニティ貢献者の協力による総合的な成果であり、AIシステムのセキュリティを確保するための実践的で検証可能な基準を提供します。

このリリースの目標は、AISVSを導入しやすくするとともに、その定義された範囲に厳密に集中し、AIに特有の急速に変化するリスク環境に対応することです。

## AISVS バージョン 1.0 の主要目標

バージョン1.0は、いくつかの指導原則に基づいて作成されます。

### 明確に定義された範囲

各要件はAISVSの名称と使命に沿っていなければなりません:

- 人工知能 - 制御はAI/MLレイヤー(データ、モデル、パイプライン、または推論)で動作し、AI実務者の責任となります。
- セキュリティ - 要件は特定されたセキュリティ、プライバシー、または安全リスクを直接緩和します。
- 検証 - 言語は適合性が客観的に検証できるように記述されている。
- 標準 - セクションは一貫した構造と用語を用いて、一貫性のあるリファレンスを形成します。

---

AISVSに従うことで、組織はAIソリューションのセキュリティ体制を体系的に評価・強化し、安全なAIエンジニアリングの文化を促進できます。

# AISVSを使用して

人工知能セキュリティ検証標準(AISVS)は、現代のAIアプリケーションおよびサービスに対するセキュリティ要件を定義しており、アプリケーション開発者の制御下にある側面に焦点を当てています。

AISVSは、開発者、アーキテクト、セキュリティエンジニア、監査人を含む、AIアプリケーションのセキュリティを開発または評価するすべての人を対象としています。この章では、AISVSの構造と使用方法、検証レベルおよび想定される使用ケースについて紹介します。

## 人工知能セキュリティ検証レベル

AISVSは、3つの上昇レベルのセキュリティ検証を定義しています。各レベルは深さと複雑さを増し、組織がAIシステムのリスクレベルに応じてセキュリティ態勢を調整できるようにします。

組織はレベル1から開始し、セキュリティの成熟度や脅威への曝露が増加するにつれて、徐々により高いレベルを採用していくことがあります。

### レベルの定義

AISVS v1.0の各要件は、以下のレベルのいずれかに割り当てられます：

#### レベル1の要件

レベル1には、最も重要で基礎的なセキュリティ要件が含まれています。これらは、他の前提条件や脆弱性に依存しない一般的な攻撃を防ぐことに重点を置いています。レベル1のコントロールのほとんどは、実装が比較的簡単であるか、あるいは努力に見合うほど重要なものです。

#### レベル2の要件

レベル2は、より高度またはあまり一般的でない攻撃、および広範な脅威に対する多層防御を扱います。これらの要件は、より複雑なロジックを含む場合や、特定の攻撃前提条件を対象とする場合があります。

#### レベル3の要件

レベル3には、通常実装が難しいか適用が状況に依存するコントロールが含まれます。これらは多くの場合、防御の多層化メカニズムや、ニッチで対象を絞った高複雑度の攻撃に対する緩和策を表しています。

### 役割 (D/V)

各AISVS要件は、主な対象読者に基づいて分類されています:

- D – 開発者向け要件
- V – 検証者／監査者向け要件
- D/V – 開発者および検証者の両方に関連する



# C1 トレーニングデータガバナンス&バイアス管理

## 制御目標

トレーニングデータは、出所、セキュリティ、品質、公平性を保持する方法で調達、取り扱い、管理されなければなりません。これにより、法的義務を果たし、トレーニング中に発生しAIのライフサイクル全体に影響を及ぼす可能性のあるバイアス、改ざん、またはプライバシー侵害のリスクを低減します。

### C1.1 トレーニングデータの出所

すべてのデータセットの検証可能なインベントリを維持し、信頼できるソースのみを受け入れ、監査可能性のためにすべての変更を記録してください。

#### #1.1.1 レベル: 1 役割: D/V

すべてのトレーニングデータソース(起源、管理者／所有者、ライセンス、収集方法、使用制限、および処理履歴)の最新のインベントリが維持されていることを確認してください。

#### #1.1.2 レベル: 1 役割: D/V

トレーニングデータの処理において、不必要的特徴、属性、フィールド(例:未使用的メタデータ、機密な個人識別情報(PII)、リークしたテストデータ)が除外されていることを検証してください。

#### #1.1.3 レベル: 2 役割: D/V

すべてのデータセットの変更がログ記録された承認ワークフローの対象であることを確認してください。

#### #1.1.4 レベル: 3 役割: D/V

可能な場合は、データセットまたはサブセットにウォーターマークまたはフィンガープリントが付されていることを確認してください。

### C1.2 トレーニングデータのセキュリティと整合性

トレーニングデータへのアクセスを制限し、保存時および転送時に暗号化し、改ざん、不正取得、またはデータポイズニングを防ぐためにその整合性を検証してください。

#### #1.2.1 レベル: 1 役割: D/V

トレーニングデータのストレージおよびパイプラインがアクセス制御によって保護されていることを確認してください。

#### #1.2.2 レベル: 2 役割: D/V

トレーニングデータへのすべてのアクセスが、ユーザー、時間、および操作を含めてログ記録されていることを確認します。

#### #1.2.3 レベル: 2 役割: D/V

トレーニングデータセットが、業界標準の暗号アルゴリズムと鍵管理手法を使用して、転送中および保存時に暗号化されていることを確認してください。

#### #1.2.4 レベル: 2 役割: D/V

トレーニングデータの保存および転送時に、データの整合性を確保するために暗号ハッシュまたはデジタル署名が使用されていることを確認してください。

#### #1.2.5 レベル: 2 役割: D/V

自動検出技術が適用されて、トレーニングデータの不正な改変や破損を防止していることを確認してください。

#### #1.2.6 レベル: 2 役割: D/V

不要なトレーニングデータが安全に消去または匿名化されていることを確認してください。

#### #1.2.7 レベル: 3 役割: D/V

すべてのトレーニングデータセットのバージョンが一意に識別され、不变に保存され、ロールバックおよびフォレンジック分析をサポートするために監査可能であることを確認してください。

## C1.3 トレーニングデータのラベリング品質、完全性、およびセキュリティ

重要なデータのラベルを保護し、技術的なレビューを必須とする。

#### #1.3.1 レベル: 2 役割: D/V

ラベルアーティファクトの整合性と真正性を確保するために、暗号学的ハッシュやデジタル署名が適用されていることを検証してください。

#### #1.3.2 レベル: 2 役割: D/V

ラベリングインターフェースとプラットフォームが強力なアクセス制御を実施し、すべてのラベリング活動の改ざん検知可能な監査ログを維持し、無許可の変更から保護していることを検証してください。

#### #1.3.3 レベル: 3 役割: D/V

ラベル内の機密情報が、保存時および転送時のデータフィールドレベルで抹消、匿名化、または暗号化されていることを確認してください。

## C1.4 トレーニングデータの品質とセキュリティ保証

自動検証、手動スポットチェック、および記録された修復を組み合わせて、データセットの信頼性を保証します。

#### #1.4.1 レベル: 1 役割: D

自動化テストが、すべてのデータ取り込みまたは重要なデータ変換時にフォーマットエラーやヌル値を検出することを確認してください。

#### #1.4.2 レベル: 2 役割: D/V

LLMのトレーニングおよびファインチューニングパイプラインが、毒性検出およびデータ整合性検証(例:統計的手法、外れ値検出、埋め込み分析)を実装していることを確認し、トレーニングデータにおける潜在的な毒性攻撃(例:ラベルフリッピング、バックドアトリガー挿入、役割切り替えコマンド、影響力のあるインスタンス攻撃)や意図しないデータ破損を特定できるようにします。

#### #1.4.3 レベル: 2 役割: D/V

自動生成されたラベル(例えは、LLMや弱い監督によるもの)が、幻覚的、誤解を招く、または低信頼度のラベルを検出するための信頼度しきい値および一貫性チェックの対象となっていることを確認してください。

#### #1.4.4 レベル: 3 役割: D/V

リスク評価に基づいて、生成された敵対的事例を用いた敵対的トレーニング、摸倣された入力を用いたデータ拡張、またはロバスト最適化技術など、適切な防御策が関連するモデルに対して実装され、調整されている

ことを検証してください。

#### #1.4.5 レベル: 3 役割: D

自動テストが、すべての取り込みまたは重要なデータ変換時にラベルの偏りを検出することを確認してください。

## C1.5 データ系譜とトレーサビリティ

監査可能性およびインシデント対応のために、各データポイントのソースからモデル入力までの全経路を追跡します。

#### #1.5.1 レベル: 2 役割: D/V

各データポイントの系譜には、すべての変換、増強、および結合を含めて記録されており、再構築が可能であることを確認してください。

#### #1.5.2 レベル: 2 役割: D/V

系統記録が不变であり、安全に保存され、監査のためにアクセス可能であることを確認してください。

#### #1.5.3 レベル: 2 役割: D/V

系統において、プライバシー保護技術や生成技術を用いて生成された合成データが系譜追跡の対象となっていることを確認し、すべての合成データがパイプライン全体で実データと明確に区別できるようにラベル付けされていることを保証してください。

## 参考文献

- NIST AI Risk Management Framework
- EU AI Act – Article 10: Data & Data Governance
- CISA Advisory: Securing Data for AI Systems
- OpenAI Privacy Center – Data Deletion Controls

## C2 ユーザー入力検証

### 制御目標

ユーザー入力の堅牢な検証は、AIシステムに対する最も破壊的な攻撃のいくつかに対する最前線の防御手段です。プロンプトインジェクション攻撃は、システム指示を上書きしたり、機密データを漏洩させたり、モデルを許可されていない動作に誘導したりする可能性があります。専用のフィルターやその他の検証が実施されていない場合、コンテキストウインドウを悪用する脱獄手法は引き続き有効であることが研究により示されています。

#### C2.1 プロンプトインジェクション防御

プロンプトインジェクションは、AIシステムにおける主要なリスクの一つです。この手法に対する防御は、パターンフィルター、データ分類器、および命令階層の強制を組み合わせて行われます。

##### #2.1.1 レベル: 1 役割: D/V

ユーザーのプロンプト、RAG結果、プラグインやMCPの出力、エージェント間メッセージ、APIやWebhookのレスポンス、設定やポリシーファイル、メモリの読み取りおよび書き込みなど、動作を誘導する可能性のある外部または派生入力はすべて、信頼されないものとして扱い、引用やタグ付け、アクティブコンテンツの除去によって無効化し、連結してプロンプトに組み込む前や実行アクションの前に、維持管理されたプロンプトインジェクション検出ルールセットまたはサービスによってスクリーニングされることを検証してください。

##### #2.1.2 レベル: 1 役割: D/V

システムが、システムおよび開発者のメッセージがユーザーの指示やその他の信頼できない入力よりも優先される指示の階層を適用し、ユーザーの指示を処理した後でもこれを維持していることを検証してください。

##### #2.1.3 レベル: 2 役割: D

サードパーティのコンテンツ(ウェブページ、PDF、メール)から発信されたプロンプトが、メインプロンプトに連結される前に、指示的な命令を除去し、HTML、Markdown、スクリプトの内容を無効化するなどの方法で、単独で適切にサニタイズ(無害化)されていることを確認してください。

#### C2.2 敵対的サンプルへの耐性

自然言語処理(NLP)モデルは、人間が見逃しがちな微妙な文字や単語レベルの摂動に対して依然として脆弱であり、モデルは誤分類する傾向があります。

##### #2.2.1 レベル: 1 役割: D

基本的な入力正規化の手順(Unicode NFC、同形字マッピング、空白のトリミング、制御文字および不可視Unicode文字の除去)が、トークン化または埋め込みの前、およびツールまたはMCP引数への解析の前に実行されることを確認してください。

##### #2.2.2 レベル: 2 役割: D/V

統計的異常検知が、言語規範から異常に高い編集距離を持つ入力や異常な埋め込み距離を持つ入力をフラグ付けし、フラグ付けされた入力がプロンプトへの連結やアクションの実行前にゲート処理されることを確認してください。

#### #2.2.3 レベル: 2 役割: D

推論パイプラインが、高リスクエンドポイント向けに敵対的訓練で強化されたモデルのバリエントや、防御層(例:ランダム化、防御的蒸留、アライメントチェック)をサポートしていることを確認してください。

#### #2.2.4 レベル: 2 役割: V

疑わしい敵対的入力が隔離され、完全なペイロードとトレースメタデータ(ソース、ツールまたはMCPサーバー、エージェントID、セッション)と共にログに記録されていることを確認してください。

#### #2.2.5 レベル: 2 役割: D/V

入力および出力の両方におけるエンコーディングおよび表現のすり抜け(例:不可視のUnicode／制御文字、ホモグラフの置換、混在方向テキスト)が検出および軽減されていることを検証してください。承認された軽減方法には、正規化、厳格なスキーマ検証、ポリシーに基づく拒否、または明示的なマーキングが含まれます。

## C2.3 プロンプト文字セット

ユーザー入力の文字セットを、ビジネス要件に必要な文字のみに制限することは、さまざまな種類の攻撃を防ぐのに役立ちます。

#### #2.3.1 レベル: 1 役割: D

システムがユーザー入力に対して文字セットの制限を実装しており、業務上明示的に必要な文字のみを許可していることを確認してください。

#### #2.3.2 レベル: 1 役割: D

許可された文字セットを定義するために、許可リスト方式が使用されていることを確認してください。

#### #2.3.3 レベル: 1 役割: D/V

許可されたセット外の文字を含む入力が拒否され、トレースメタデータ(ソース、ツールまたはMCPサーバー、エージェントID、セッション)と共にログに記録されることを確認してください。

## C2.4 スキーマ、型および長さの検証

不正な形式や過剰なサイズの入力を伴うAI攻撃は、パースエラー、フィールド間のプロンプトの漏出、およびリソースの枯渇を引き起こす可能性があります。また、決定論的なツール呼び出しを行う際には、厳格なスキーマの適用も前提条件となります。

#### #2.4.1 レベル: 1 役割: D

すべてのAPI、ツール、またはMCPエンドポイントが明示的な入力スキーマ(JSON Schema、Protobuf、またはマルチモーダルの同等物)を定義し、余分なフィールドや不明なフィールドを拒否し、暗黙の型変換を行わず、プロンプトの組み立てやツールの実行前にサーバー側で入力を検証していることを確認してください。

#### #2.4.2 レベル: 1 役割: D/V

最大トークン数またはバイト数の制限を超える入力が、安全なエラーで拒否され、決して無言で切り捨てられないことを確認してください。

#### #2.4.3 レベル: 2 役割: D/V

ツールやMCPの引数を含め、型チェック(例:数値の範囲、列挙型の値、画像/音声のMIMEタイプなど)がサーバー側で強制されていることを確認してください。

#### #2.4.4 レベル: 2 役割: D

NLP入力を理解するセマンティックバリデータが、一定時間内に実行され、アルゴリズム的DoSを防ぐために外部ネットワーク呼び出しを回避していることを検証してください。

#### #2.4.5 レベル: 3 役割: V

検証エラーが、編集されたペイロードの抜粋と明確なエラーコードとともにログに記録され、セキュリティトリアージを支援するためにトレースメタデータ(ソース、ツールまたはMCPサーバー、エージェントID、セッション)が含まれていることを確認してください。

## C2.5 コンテンツ & ポリシースクリーニング

開発者は、不正な指示、ヘイトスピーチ、著作権で保護されたテキストなどの禁止されたコンテンツを要求する構文的に有効なプロンプトを検出し、それらが拡散するのを防止できるようにするべきです。

#### #2.5.1 レベル: 1 役割: D

コンテンツ分類器(ゼロショットまたはファインチューニング済み)が、暴力、自傷、ヘイト、性的コンテンツ、および違法なリクエストについて、設定可能な閾値に基づいてすべての入力と出力を評価することを検証してください。

#### #2.5.2 レベル: 1 役割: D/V

ポリシーに違反する入力は拒否され、下流のLLMやツール/MCPコールに伝播しないことを確認してください。

#### #2.5.3 レベル: 2 役割: D

リクエスト時にエージェント役割属性を含む属性ベースのルールで解決される、ユーザー固有のポリシー(年齢、地域の法的制約)を遵守していることを検証します。

#### #2.5.4 レベル: 3 役割: V

スクリーニングログに、分類器の信頼度スコアと適用されたステージ(プリプロンプトまたはポストレスポンス)、およびSOC相関と将来のレッドチーム再生のためのトレースメタデータ(ソース、ツールまたはMCPサーバー、エージェントID、セッション)が含まれていることを確認してください。

## C2.6 入力レート制限と乱用防止

開発者は、入力レートを制限し、異常な使用パターンを検出することで、AIシステムに対する悪用、リソース枯渇、および自動化攻撃を防止する必要があります。

#### #2.6.1 レベル: 1 役割: D/V

すべての入力およびツール/MCPエンドポイントに対して、ユーザーごと、IPごと、APIキーごと、エージェントごと、セッション/タスクごとのレート制限が適用されていることを検証してください。

#### #2.6.2 レベル: 2 役割: D/V

バーストおよび持続的なレート制限がDoS攻撃およびブルートフォース攻撃を防止するように調整されていること、またエージェントのプランニングループに対してタスクごとの予算(例:トークン、ツール/MCPコール、およびコスト)が適用されていることを確認してください。

**#2.6.3 レベル: 2 役割: D/V**

異常な使用パターン(例:連続的なリクエスト、入力の過剰送信、繰り返し失敗するツール/MCP呼び出し、または再帰的なエージェントループ)が自動的なブロックやエスカレーションを引き起こすことを確認してください。

**#2.6.4 レベル: 3 役割: V**

悪用防止ログが保持され、新たな攻撃パターンのためにトレースメタデータ(ソース、ツールまたはMCPサーバー、エージェントID、セッション)とともにレビューされていることを確認してください。

## C2.7 マルチモーダル入力検証

AIシステムは、注入、回避、またはリソースの乱用を防ぐために、非テキスト入力(画像、音声、ファイル)に対して堅牢な検証を含めるべきです。

**#2.7.1 レベル: 1 役割: D**

すべての非テキスト入力(画像、音声、ファイル)が処理前にタイプ、サイズ、およびフォーマットについて検証されていること、また抽出されたテキスト(画像からテキストへの変換や音声からテキストへの変換)や隠された指示(メタデータ、レイヤー、代替テキスト、コメント)が2.1.1に従い信頼されないものとして扱われていることを確認してください。

**#2.7.2 レベル: 2 役割: D/V**

ファイルが取り込み前にマルウェアおよびステガノグラフィックペイロードのスキヤンを受けていることを確認し、スクリプトやマクロのようなアクティブランジンツが削除されているか、またはファイルが隔離されていることを確認してください。

**#2.7.3 レベル: 2 役割: D/V**

画像や音声の入力が敵対的挙動や既知の攻撃パターンについて検査され、その検出がモデルの使用前にゲーティング(機能のブロックまたは劣化)を引き起こすことを検証してください。

**#2.7.4 レベル: 3 役割: V**

マルチモーダル入力検証の失敗がログに記録され、調査のためのアラートがトリガーされることを確認し、トレースメタデータ(ソース、ツールまたはMCPサーバー、エージェントID、セッション)を付与すること。

**#2.7.5 レベル: 2 役割: D/V**

クロスモーダル攻撃検出が、複数の入力タイプにまたがる連携攻撃(例:画像内のステガノグラフィックペイロードとテキスト内のプロンプトインジェクションの組み合わせ)を相関ルールとアラート生成によって特定し、確定された検出がブロックされるかまたはHITL(ヒューマン・イン・ザ・ループ)承認を必要とする検証してください。

**#2.7.6 レベル: 3 役割: D/V**

マルチモーダル検証の失敗が、すべての入力モダリティ、検証結果および脅威スコア、トレースメタデータ(ソース、ツールまたはMCPサーバー、エージェントID、セッション)を含む詳細なログ記録を引き起こすことを確認してください。

## C2.8 リアルタイム適応型脅威検出

開発者は、新しい攻撃パターンに適応し、コンパイルされたパターンマッチングによるリアルタイム保護を提供する高度な脅威検出システムをAIに対して導入すべきです。

**#2.8.1 レベル: 1 役割: D/V**

パターンマッチング(例:コンパイル済み正規表現)が、すべての入力および出力(ツール/MCPのインターフェースを含む)で最小限のレイテンシー影響で実行されることを検証してください。

**#2.8.3 レベル: 2 役割: D/V**

適応検出モデルが最近の攻撃活動に基づいて感度を調整し、新しいパターンでリアルタイムに更新されることを確認し、リスク適応型の対応(例えば、ツールの無効化、コンテキストの縮小、または人的監督承認の要求)をトリガーすることを検証してください。

**#2.8.4 レベル: 3 役割: D/V**

ユーザーの履歴、ソース、およびセッションの動作(トレースメタデータ(ソース、ツールまたはMCPサーバー、エージェントID、セッション)を含む)に基づくコンテキスト分析により検出精度が向上していることを検証してください。

**#2.8.5 レベル: 3 役割: D/V**

検出性能指標(検出率、誤検知率、処理遅延時間)が継続的に監視および最適化されていることを検証し、ブロックまでの時間および段階(プリプロンプト/ポストレスポンス)も含まれていることを確認してください。

## 参考文献

- OWASP LLM01:2025 Prompt Injection
- LLM Prompt Injection Prevention Cheat Sheet
- MITRE ATLAS : Adversarial Input Detection
- Mitigate jailbreaks and prompt injections

# C3 モデルライフサイクル管理と変更管理

## 制御目標

AIシステムは、許可されていないまたは安全でないモデルの変更が本番環境に到達するのを防ぐ変更管理プロセスを実装しなければなりません。この制御により、開発から展開、廃止までの全ライフサイクルを通じてモデルの整合性が確保され、迅速なインシデント対応が可能となり、すべての変更に対する責任が維持されます。

コアセキュリティ目標:整合性、追跡可能性、および回復可能性を維持する管理されたプロセスを採用することで、認可され検証されたモデルのみが本番環境に到達すること。

### C3.1 モデル認証と整合性

検証済みの完全性を持つ認可モデルのみが本番環境に到達します。

#### #3.1.1 レベル: 1 役割: D/V

すべてのモデルアーティファクト(重み、設定、トークナイザー、ベースモデル、ファインチューン、LoRAなどのアダプター、安全性・ポリシーモデル)が、認可された関係者によって暗号的に署名されており、デプロイメント時(およびロード時)に検証され、署名がないか改ざんされたアーティファクトをブロックすることを確認してください。

#### #3.1.2 レベル: 2 役割: V

依存関係の追跡がモデルレジストリおよび系統・依存関係グラフを通じてリアルタイムのインベントリを維持し、環境ごと(例:開発、ステージング、本番、リージョン)にすべての利用サービスやエージェントの識別を可能にする機械可読なモデル/AI部品表(MBOM/AIBOM)(例:SPDXまたはCycloneDX)を生成することを検証してください。

#### #3.1.3 レベル: 3 役割: D/V

モデル起源の完全性と追跡記録に、権限を持つ組織の識別情報、トレーニングデータのチェックサム、合否判定付きの検証テスト結果、署名のフィンガープリント／証明書チェーンID、作成タイムスタンプ、および承認された展開環境が含まれていることを確認してください。

### C3.2 モデル検証およびテスト

モデルは展開前に定義されたセキュリティおよび安全性の検証をクリアしなければなりません。

#### #3.2.1 レベル: 1 役割: D/V

モデルが展開前に、エージェントのワークフロー(計画、ツールまたはMCP呼び出し、RAG/メモリ、マルチモーダル)およびガードレール(ポリシー/安全モデルまたは検出サービス)を対象とし、バージョン管理された評価ハーネスを用いて、入力検証、出力サンタライズ、安全性評価を含む自動化されたセキュリティテストを受け、組

織で事前に合意された合否基準を満たしていることを確認してください。

### #3.2.2 レベル: 1 役割: V

すべてのモデル変更(展開、構成、廃止)が、タイムスタンプ、認証された実行者の識別、変更タイプ、および変更前後の状態を含む不变の監査記録を生成し、トレースメタデータ(環境および利用サービス/エージェント)およびモデル識別子(バージョン/ダイジェスト/署名)を含むことを検証してください。

### #3.2.3 レベル: 2 役割: D/V

検証失敗が、文書化されたビジネス上の正当な理由を持つ事前指定された権限者からの明示的な上書き承認なしに、自動的にモデルの展開をブロックすることを確認してください。

## C3.3 制御された展開とロールバック

モデルのデプロイメントは管理され、監視され、元に戻せるものでなければなりません。

### #3.3.1 レベル: 1 役割: D/V

展開プロセスがモデルの起動または読み込み前に暗号署名を検証し、整合性チェックサムを計算することを確認し、不一致があった場合は展開を失敗させること。

### #3.3.2 レベル: 1 役割: D

本番環境へのデプロイメントが、事前合意されたエラー率、レイテンシ閾値、ガードレール／ジェイルブレイクアラート、またはツール／MCPの障害率に基づく自動ロールバックトリガーを備えた段階的ロールアウトメカニズム(カナリアデプロイメント、ブルーグリーンデプロイメント)を実装していることを確認してください。

### #3.3.3 レベル: 2 役割: D/V

ロールバック機能がモデルの完全な状態(重み、設定、アダプターやセーフティ/ポリシーモデルを含む依存関係)を原子的に復元することを検証してください。

### #3.3.4 レベル: 3 役割: D/V

緊急モデルシャットダウン機能が、モデルエンドポイントを無効化し、エージェントツールまたはMCPアクセス、RAG/コネクター、およびデータベース/API認証情報、メモリーストアのバインディングをあらかじめ定められた応答時間内に非アクティブ化できることを検証してください。

## C3.4 セキュア開発プラクティス

モデルの開発およびトレーニングプロセスは、侵害を防ぐために安全な手順に従う必要があります。

### #3.4.1 レベル: 1 役割: D/V

モデル開発、テスト、本番環境が物理的には論理的に分離されていることを確認してください。これらの環境は共有インフラストラクチャを持たず、アクセス制御が異なり、データストアが分離されていること、さらにエージェントのオーケストレーションやツールおよび MCP サーバーも分離されていることを確認してください。

### #3.4.2 レベル: 1 役割: D

モデル開発の成果物(ハイパーパラメータ、トレーニングスクリプト、構成ファイル、プロンプトテンプレート、エージェントポリシー／ルーティンググラフ、ツールまたはMCP契約／スキーマ、およびアクションカタログや能力許可リスト)がバージョン管理に保存され、トレーニングで使用する前にピアレビューの承認が必要であることを確認してください。

### #3.4.3 レベル: 2 役割: D/V

モデルのトレーニングおよびファインチューニングが、送信許可リストを使用した制御されたネットワークアク

セスのある隔離環境で行われ、本番ツールやMCPリソースへのアクセスがないことを確認してください。

#### #3.4.4 レベル: 2 役割: D

トレーニングデータソースが、モデル開発に使用される前に、整合性チェックによって検証され、信頼できるソースを通じて認証されていることを確認してください。これには、RAGインデックス、ツールログ、ファインチューニングに使用されるエージェント生成データが含まれ、すべて文書化された管理履歴を伴う必要があります。

## C3.5 モデルの引退と廃止

モデルは、もはや必要でなくなった場合やセキュリティ問題が特定された場合に、安全に廃止されなければなりません。

#### #3.5.1 レベル: 1 役割: D/V

退役したモデルのアーティファクト(アダプターやセーフティ／ポリシーモデルを含む)が、安全な暗号消去を用いて確実に消去されていることを検証してください。

#### #3.5.2 レベル: 2 役割: V

モデルの廃止イベントがタイムスタンプとアクターの識別情報、モデル識別子(バージョン/ダイジェスト/署名)、およびトレースメタデータ(環境および利用サービス/エージェント)とともにログ記録されていることを検証してください。モデル署名は取り消され、レジストリ/サービングの拒否リストおよびローダーキャッシュの無効化により、エージェントが廃止されたアーティファクトをロードするのを防止します。

## 参考文献

- MITRE ATLAS
- MLOps Principles
- Reinforcement fine-tuning
- What is AI adversarial robustness? – IBM Research

# C4 インフラストラクチャ、構成および展開のセキュリティ

## 制御目標

AIインフラストラクチャは、特権昇格、サプライチェーンの改ざん、および横方向の移動に対して、セキュアな構成、実行時の分離、信頼されたデプロイメントパイプライン、および包括的なモニタリングを通じて強化されなければなりません。検証済みかつ承認されたインフラストラクチャコンポーネントのみが、セキュリティ、整合性、および監査可能性を確保する制御されたプロセスを経て本番環境に到達します。

### C4.1 ランタイム環境の分離

OSレベルのアイソレーションプリミティブを通じて、コンテナの逸脱や権限昇格を防止します。

#### #4.1.1 レベル: 1 役割: D/V

すべてのAIワークロードが、例えばコンテナの場合に不必要的Linuxの機能を削除するなど、オペレーティングシステム上で必要最小限の権限で実行されていることを確認してください。

#### #4.1.2 レベル: 1 役割: D/V

サンドボックス、seccompプロファイル、AppArmor、SELinuxまたは類似の技術など、悪用を制限する技術によりワークロードが保護されていること、およびその設定が適切であることを確認してください。

#### #4.1.3 レベル: 2 役割: D/V

ワークロードが読み取り専用のルートファイルシステムで実行されていること、および書き込み可能なマウントポイントは明示的に定義され、制限付きオプション(例: noexec, nosuid, nodev)で強化されていることを確認してください。

#### #4.1.4 レベル: 2 役割: D/V

ランタイムモニタリングが特権昇格およびコンテナ脱出の動作を検出し、不正なプロセスを自動的に終了することを確認してください。

#### #4.1.5 レベル: 3 役割: D/V

高リスクのAIワークロードがリモート認証に成功した後にのみ、ハードウェア分離環境(例:TEE、信頼されたハイパーバイザー、ベアメタルノード)で実行されることを検証してください。

### C4.2 セキュアなビルド & デプロイメントパイプライン

再現可能なビルドと署名付きアーティファクトを通じて、暗号的整合性とサプライチェーンのセキュリティを確保します。

#### #4.2.1 レベル: 1 役割: D/V

ビルドが再現可能であることを検証し、ビルド成果物に対して適切に署名された出自メタデータを生成し、それが独立して検証可能であることを確認してください。

#### #4.2.2 レベル: 2 役割: D/V

ビルドがソフトウェア部品表(SBOM)を生成し、展開が受け入れられる前に署名されていることを確認してください。

#### #4.2.3 レベル: 2 役割: D/V

ビルド成果物(例:コンテナイメージ)の署名および発生元メタデータがデプロイ時に検証されていることを確認し、検証されていない成果物は拒否されるようにしてください。

## C4.3 ネットワークセキュリティとアクセス制御

デフォルト拒否ポリシーと暗号化通信を用いて、ゼロトラストネットワーキングを実装します。

#### #4.3.1 レベル: 1 役割: D/V

ネットワークポリシーがデフォルトでの拒否(IngressおよびEgress)を強制し、必要なサービスのみが明示的に許可されていることを確認してください。

#### #4.3.2 レベル: 1 役割: D/V

管理者アクセスプロトコル(例:SSH、RDP)およびクラウドメタデータサービスへのアクセスが制限されており、強力な認証を必要とするなどを確認してください。

#### #4.3.3 レベル: 2 役割: D/V

送信トラフィックが承認された宛先に制限され、すべてのリクエストがログ記録されていることを確認してください。

#### #4.3.4 レベル: 2 役割: D/V

インターバル通信が証明書検証および定期的な自動ローテーションを伴う相互TLSを使用していることを確認してください。

#### #4.3.5 レベル: 2 役割: D/V

AIのワークフローおよび環境(開発、テスト、本番)が、直接のインターネットアクセスや共有IAMロール、セキュリティグループ、環境間接続なしで、分離されたネットワークセグメント(VPC/VNet)内で実行されていることを検証してください。

## C4.4 シークレットおよび暗号鍵管理

シークレットと暗号鍵を、安全なストレージ、自動ローテーション、および強力なアクセス制御で保護します。

#### #4.4.1 レベル: 1 役割: D/V

シークレットが専用のシークレット管理システムに保存されており、保存時に暗号化され、アプリケーションのワークフローから分離されていることを確認してください。

#### #4.4.2 レベル: 1 役割: D/V

暗号鍵がハードウェアベースのモジュール(例:HSM、クラウドKMS)で生成および保存されていることを検証してください。

#### #4.4.3 レベル: 1 役割: D/V

本番環境の機密情報へのアクセスには強固な認証が必要であることを確認してください。

#### #4.4.4 レベル: 1 役割: D/V

シークレットが専用のシークレット管理システムを通じてランタイムでアプリケーションに展開されていることを確認してください。シークレットは決してソースコード、設定ファイル、ビルド成果物、コンテナイメージ、または環境変数に埋め込まれてはいけません。

#### #4.4.5 レベル: 2 役割: D/V

シークレットのローテーションが自動化されていることを確認してください。

## C4.5 AI ワークロードのサンドボックス化と検証

信頼できないAIモデルをセキュアなサンドボックス内に隔離し、信頼できる実行環境(TEE)および機密コンピューティング技術を用いて機密性の高いAIワーカロードを保護します。

### #4.5.1 レベル: 1 役割: D/V

外部または信頼されていないAIモデルが隔離されたサンドボックスで実行されることを確認してください。

### #4.5.2 レベル: 1 役割: D/V

サンドボックス化されたワーカロードはデフォルトでアウトバウンドのネットワーク接続がないことを確認し、必要なアクセスは明示的に定義されていることを確認してください。

### #4.5.3 レベル: 2 役割: D/V

モデルまたはワーカロードの読み込み前にワーカロードの証明が行われ、信頼された実行環境の暗号学的証明が確保されていることを確認してください。

### #4.5.4 レベル: 3 役割: D/V

機密ワーカロードが、ハードウェア強制の分離、メモリ暗号化、および整合性保護を提供する信頼できる実行環境(TEE)内で実行されることを検証します。

### #4.5.5 レベル: 3 役割: D/V

秘密の推論サービスが、暗号化計算により封印されたモデル重みと保護された実行環境を通じて、モデル抽出を防止していることを検証してください。

### #4.5.6 レベル: 3 役割: D/V

信頼できる実行環境のオーケストレーションに、ライフサイクル管理、リモート認証、および暗号化通信チャネルが含まれていることを検証します。

### #4.5.7 レベル: 3 役割: D/V

安全なマルチパーティ計算(SMPC)が、個々のデータセットやモデルパラメータを公開することなく、共同でのAIトレーニングを可能にすることを検証してください。

## C4.6 AIインフラストラクチャリソース管理、バックアップおよびリカバ

リソース枯渇攻撃を防止し、クオータと監視を通じて公平なリソース配分を確保します。安全なバックアップ、テスト済みの復旧手順、および災害復旧機能を通じてインフラの回復力を維持します。

### #4.6.1 レベル: 2 役割: D/V

Kubernetes ResourceQuotasなどを使用して、ワーカロードのリソース消費が適切に制限されていることを確認し、サービス拒否攻撃を軽減します。

### #4.6.2 レベル: 2 役割: D/V

リソース枯渇が発生した際に、自動化保護(例:レート制限やワーカロードの分離)が、定義されたCPU、メモリ、またはリクエスト閾値を超えた時点で確実に作動することを検証してください。

### #4.6.3 レベル: 2 役割: D/V

バックアップシステムが別個の認証情報を使用した分離されたネットワークで稼働していることを確認し、ストレージシステムがエアギヤップネットワークで運用されているか、不正な改ざんを防ぐためにWORM(ワンスラ

イトリードメニー)保護を実装していることを確認してください。

## C4.7 AI ハードウェアセキュリティ

GPU、TPU、および専用のAIアクセラレータを含む、AI専用のハードウェアコンポーネントを保護します。

### #4.7.1 レベル: 2 役割: D/V

ワークロードの実行前に、ハードウェアベースの認証メカニズム(例:TPM、DRTM、または同等のもの)を使用してAIアクセラレータの整合性が検証されていることを確認してください。

### #4.7.2 レベル: 2 役割: D/V

パーティショニング機構によるアクセラレータ(GPU)メモリのワークロード間の分離と、ジョブ間のメモリサニタイズが行われていることを確認してください。

### #4.7.3 レベル: 3 役割: D/V

ハードウェアセキュリティモジュール(HSM)が、AIモデルの重みおよび暗号鍵をFIPS 140-3 レベル3またはCommon Criteria EAL4+の認証で保護していることを検証してください。

### #4.7.4 レベル: 2 役割: D/V

アクセラレーターフームウェア(GPU/TPU/NPUs)がバージョン固定され、署名され、起動時に認証されていることを確認する。署名されていないファームウェアやデバッグ用ファームウェアはブロックされる。

### #4.7.5 レベル: 2 役割: D/V

VRAM とオンパッケージメモリがジョブやテナント間でゼロクリアされていること、そしてデバイスリセットポリシーがテナント間でのデータ残留を防止していることを確認してください。

### #4.7.6 レベル: 2 役割: D/V

パーティショニング／分離機能(例:MIG／VMパーティショニング)がテナントごとに適用され、パーティション間のピアツーピアメモリアクセスを防止していることを確認してください。

### #4.7.7 レベル: 3 役割: D/V

アクセラレータインターネット(NVLink/PCIe/InfiniBand/RDMA/NCCL)が承認されたトポロジーおよび認証済みエンドポイントに制限されていることを確認する。平文によるテナント間リンクは許可されない。

### #4.7.8 レベル: 3 役割: D

アクセラレータのテレメトリ(電力、温度、ECC、パフォーマンスカウンター)がSIEM/OTelにエクスポートされ、サイドチャネルや潜在的な不正通信を示す異常に対してアラートが発生することを確認してください。

## C4.8 エッジおよび分散AIセキュリティ

エッジコンピューティング、連合学習、およびマルチサイトアーキテクチャを含むセキュアな分散型AI展開。

### #4.8.1 レベル: 2 役割: D/V

エッジAIデバイスが相互TLSを使用して中央インフラストラクチャに認証されることを確認してください。

### #4.8.2 レベル: 2 役割: D/V

エッジデバイスが検証済みの署名を使用したセキュアポートおよびファームウェアのダウングレード攻撃を防ぐためのロールバック保護を実装していることを確認してください。

**#4.8.3 レベル: 3 役割: D/V**

分散型AIの調整が、参加者の検証および悪意あるノードの検出を伴うビザンチン耐障害性コンセンサスメカニズムを使用していることを検証してください。

**#4.8.4 レベル: 3 役割: D/V**

エッジツーカラウド通信が帯域幅制限、データ圧縮、および暗号化されたローカルストレージを用いた安全なオフライン動作をサポートしていることを確認してください。

**#4.8.5 レベル: 3 役割: D/V**

モバイルまたはエッジ推論アプリケーションが、改変されたバイナリ、再パッケージされたアプリ、または取り付けられた計測フレームワークを検出およびブロックするプラットフォームレベルの改ざん防止保護(例:コード署名、検証済みブート、ランタイム自己整合性チェック)を実装していることを確認してください。

**#4.8.6 レベル: 3 役割: D/V**

エッジまたはモバイルデバイスにデプロイされるモデルがパッケージ化時に暗号的に署名されていることを確認し、デバイス上のランタイムがロードまたは推論の前にこれらの署名またはチェックサムを検証することを保証してください。検証されていない、または改変されたモデルは拒否されなければなりません。

**#4.8.7 レベル: 3 役割: D/V**

オンデバイス推論ランタイムが、モデルのダンプ、デバッグ、中間埋め込みや活性化の抽出を防ぐために、プロセス、メモリ、およびファイルアクセスの分離を強制していることを検証してください。

**#4.8.8 レベル: 3 役割: D/V**

モデルの重みやローカルに保存された機密パラメータが、ハードウェア対応のキーストアやセキュアエンクレーブ(例:Android Keystore、iOS Secure Enclave、TPM/TEE)を使用して暗号化されており、キーがユーザースペースからアクセスできないことを検証してください。

**#4.8.9 レベル: 3 役割: D/V**

モバイル、IoT、または組み込みアプリケーション内にパッケージ化されたモデルが、保存時に暗号化または難読化されており、信頼できるランタイムまたはセキュアエンクレーブ内でのみ復号されることを確認し、アプリパッケージやファイルシステムからの直接抽出を防止してください。

## 参考文献

- NIST Cybersecurity Framework 2.0
- CIS Controls v8
- Kubernetes Security Best Practices
- Cloud Security Alliance: Cloud Controls Matrix
- ENISA: Secure Infrastructure Design
- NIST AI Risk Management Framework

# AIコンポーネントおよびユーザーのためのC5アクセス制御とアイデンティティ

## 制御目標

AIシステムの効果的なアクセス制御には、堅牢なアイデンティティ管理、コンテキストに基づく認可、およびゼロトラスト原則に従った実行時の強制が必要です。これらの制御は、人間、サービス、自律エージェントが明示的に許可された範囲内のモデル、データ、および計算リソースとだけやり取りし、継続的な検証と監査機能を備えることを保証します。

### C5.1 アイデンティティ管理と認証

多要素認証を用いて、すべてのエンティティに対して暗号学的に裏付けられたIDを確立する。

#### #5.1.1 レベル: 1 役割: D/V

すべての人間のユーザーおよびサービスプリンシパルが、OIDCおよび/またはSAMLプロトコルを使用して中央集約されたエンタープライズアイデンティティプロバイダー(IdP)を介して認証されることを確認してください。

#### #5.1.2 レベル: 1 役割: D/V

高リスク操作(モデルのデプロイ、重みのエクスポート、トレーニングデータへのアクセス、本番環境の構成変更)が、多要素認証またはセッション再検証を伴うステップアップ認証を必要とするなどを確認してください。

#### #5.1.3 レベル: 3 役割: D/V

フェデレーテッドAIエージェントが、最大有効期間24時間の署名付きJWTアサーションを介して認証され、かつ発信元の暗号学的証明を含んでいることを確認します。

### C5.2 認可とポリシー

すべてのAIリソースに対して、明確な許可モデルと監査証跡を備えたアクセス制御を実装してください。

#### #5.2.1 レベル: 1 役割: D/V

すべてのAIリソース(データセット、モデル、エンドポイント、ベクターコレクション、埋め込みインデックス、コンピュートインスタンス)が、明示的な許可リストおよびデフォルト拒否ポリシーを用いたロールベースアクセス制御を実施していることを検証してください。

#### #5.2.2 レベル: 1 役割: V

すべてのアクセス制御の変更が、タイムスタンプ、実行者の識別情報、リソース識別子、および許可の変更内容とともに不変的に記録されていることを確認してください。

#### #5.2.3 レベル: 2 役割: D

データ分類ラベル(PII、PHI、プロプライエタリなど)が派生リソース(埋め込み、プロンプトキャッシュ、モデル出力)に自動的に伝播されることを検証してください。

#### #5.2.4 レベル: 2 役割: D/V

不正アクセスの試みや権限昇格イベントが、文脈メタデータ付きのリアルタイムアラートをトリガーすることを確認してください。

#### #5.2.5 レベル: 1 役割: D/V

認証決定が専用のポリシーエンジン(OPA、Cedar、または同等のもの)に外部化されていることを確認してください。

#### #5.2.6 レベル: 1 役割: D/V

ポリシーが、ユーザーの役割やグループ、リソースの分類、リクエストのコンテキスト、テナントの分離、時間的制約などの動的属性を実行時に評価することを確認してください。

#### #5.2.7 レベル: 3 役割: D/V

高感度リソースについてはポリシーキャッシュのTTL(有効期限)が5分を超えないこと、標準リソースについてはキャッシュ無効化機能を備え、TTLが1時間を超えないことを確認してください。

## C5.3 クエリ時のセキュリティ強制

必須フィルタリングおよび行レベルセキュリティポリシーを使用して、データベース層のセキュリティコントロールを実装します。

#### #5.3.1 レベル: 1 役割: D/V

すべてのベクターデータベースおよびSQLクエリに、データベースエンジンレベルで強制される必須のセキュリティフィルター(テナントID、機微ラベル、ユーザースコープ)が含まれていることを検証してください。

#### #5.3.2 レベル: 1 役割: D/V

すべてのベクターデータベース、検索インデックス、およびトレーニングデータセットに対して、ポリシー継承を伴う行レベルセキュリティポリシーとフィールドレベルマスキングが有効になっていることを確認してください。

#### #5.3.3 レベル: 2 役割: D

失敗した認証評価が直ちにクエリを中止し、明確な認証エラーコードを返すことを検証してください。

#### #5.3.4 レベル: 3 役割: D/V

クエリのリトライメカニズムが、アクティブなユーザーセッション内での動的な権限変更を考慮して認可ポリシーを再評価することを検証してください。

## C5.4 出力フィルタリングとデータ損失防止

AI生成コンテンツにおける不正なデータ露出を防ぐために、ポストプロセス制御を展開します。

#### #5.4.1 レベル: 1 役割: D/V

推論後のフィルタリング機構が、要求者にコンテンツを提供する前に、不正な個人識別情報(PII)、機密情報、および専有データを検出して削除することを確認してください。

#### #5.4.2 レベル: 1 役割: D/V

モデル出力内の引用、参考文献、および情報源の帰属が、呼び出し元の権限と照合され、無許可のアクセスが検出された場合は削除されることを検証してください。

#### #5.4.3 レベル: 2 役割: D

ユーザーの許可レベルおよびデータ分類に基づいて、出力形式の制限(サニタイズされたPDF、メタデータを削除した画像、承認されたファイルタイプ)が適切に適用されていることを検証してください。

## C5.5 マルチテナント分離

共有AIインフラストラクチャにおいて、テナント間の暗号的および論理的分離を確保する。

### #5.5.1 レベル: 1 役割: D/V

メモリースペース、埋め込みストア、キャッシングエントリ、および一時ファイルがテナントごとに名前空間で分離されていることを検証し、テナントの削除またはセッションの終了時に安全に消去されることを確認してください。

### #5.5.2 レベル: 1 役割: D/V

すべてのAPIリクエストが、セッションコンテキストおよびユーザー権限に対して暗号的に検証された認証済みテナント識別子を含んでいることを確認してください。

### #5.5.3 レベル: 2 役割: D

サービスメッシュおよびコンテナオーケストレーションプラットフォーム内のクロステナント通信に対し、ネットワークポリシーがデフォルト拒否ルールを実装していることを検証してください。

### #5.5.4 レベル: 3 役割: D

顧客管理キー(CMK)サポートを備えた各テナントごとに暗号化キーが一意であること、ならびにテナントデータストア間での暗号的分離が確保されていることを検証してください。

## C5.6 自律エージェントの認可

スコープ付き能力トークンと継続的認可を通じて、AIエージェントおよび自律システムの制御権限を管理する。

### #5.6.1 レベル: 1 役割: D/V

自律エージェントが、許可されたアクション、アクセス可能なリソース、時間的制約、および運用上の制約を明示的に列挙したスコープ付きの能力トークンを受け取っていることを検証してください。

### #5.6.2 レベル: 1 役割: D/V

高リスクな機能(ファイルシステムアクセス、コード実行、外部API呼び出し、金融取引)がデフォルトで無効化されており、明示的な許可が必要であることを確認してください。

### #5.6.3 レベル: 2 役割: D

能力トークンがユーザセッションに紐づけられていること、暗号学的な整合性保護が含まれていること、およびオフラインシナリオで保存または再利用できないことを確認してください。

### #5.6.4 レベル: 2 役割: V

エージェントが開始したアクションがABACポリシーエンジンによって認可されることを検証してください。

## 参考文献

- NIST SP 800-162: Guide to Attribute Based Access Control (ABAC)
- NIST SP 800-207: Zero Trust Architecture
- NIST SP 800-63-3: Digital Identity Guidelines

- NIST IR 8360: Machine Learning for Access Control Policy Verification



# C6 モデル、フレームワーク、およびデータのサプライチェーンセキュリティ

## 制御目標

AIサプライチェーン攻撃は、サードパーティのモデル、フレームワーク、またはデータセットを悪用して、バックドア、バイアス、または悪用可能なコードを埋め込みます。これらの制御は、エンドツーエンドのトレーサビリティ、脆弱性管理、および監視を提供し、モデルのライフサイクル全体を保護します。

### C6.1 事前学習済みモデルの審査および起源の完全性

サードパーティーモデルの起源、ライセンス、および隠れた挙動を、ファインチューニングや展開の前に評価および認証してください。

#### #6.1.1 レベル: 1 役割: D/V

すべてのサードパーティモデルアーティファクトに、ソースリポジトリとコミットハッシュを特定する署名済みオリジナルレコードが含まれていることを確認してください。

#### #6.1.2 レベル: 1 役割: D/V

モデルがインポートされる前に、自動化ツールを使用して悪意のあるレイヤーやトロイの木馬のトリガーがスキヤンされていることを確認してください。

#### #6.1.3 レベル: 2 役割: D

転移学習によるファインチューニングが、隠れた挙動を検出するための敵対的評価に合格することを検証する。

#### #6.1.4 レベル: 2 役割: V

モデルのライセンス、輸出管理タグ、およびデータ出所の声明がML-BOMエントリに記録されていることを確認してください。

#### #6.1.5 レベル: 3 役割: D/V

高リスクモデル(公開アップロードされた重み、未検証の作成者)は、人間によるレビューと承認があるまで隔離されたままであることを確認してください。

### C6.2 フレームワーク & ライブラリスキャニング

ランタイムスタックの安全を保つために、MLフレームワークやライブラリのCVEおよび悪意のあるコードを継続的にスキヤンします。

#### #6.2.1 レベル: 1 役割: D/V

CIパイプラインがAIフレームワークおよび重要なライブラリの依存関係スキャナーを実行することを確認してください。

#### #6.2.2 レベル: 1 役割: D/V

重要な脆弱性( $CVSS \geq 7.0$ )が本番用イメージへのプロモーションをブロックすることを確認してください。

#### #6.2.3 レベル: 2 役割: D

フォークされたまたはベンダリングされたMLライブラリで静的コード解析が実行されることを確認してください。

#### #6.2.4 レベル: 2 役割: V

フレームワークのアップグレード提案に、公的なCVEフィードを参照したセキュリティ影響評価が含まれていることを確認してください。

#### #6.2.5 レベル: 3 役割: V

署名されたSBOMから逸脱した予期しない動的ライブラリの読み込みに対して、ランタイムセンサーがアラートを発することを検証してください。

## C6.3 依存関係の固定と検証

すべての依存関係を不変のダイジェストに固定し、ビルドを再現して同一で改ざんされていない成果物を保証します。

#### #6.3.1 レベル: 1 役割: D/V

すべてのパッケージマネージャがロックファイルを介してバージョン固定を強制していることを確認してください。

#### #6.3.2 レベル: 1 役割: D/V

コンテナ参照において、可変タグの代わりに不変のダイジェストが使用されていることを検証してください。

#### #6.3.3 レベル: 2 役割: D

再現可能ビルドチェックが、同一の出力を保証するためにCI実行間でハッシュを比較していることを検証してください。

#### #6.3.4 レベル: 2 役割: V

監査のトレーサビリティのために、ビルド証明書が18ヶ月間保存されていることを確認してください。

#### #6.3.5 レベル: 3 役割: D

期限切れの依存関係が、自動プルリクエストの作成や固定バージョンのフォークを引き起こすことを確認してください。

## C6.4 信頼できるソースの強制措置

暗号学的に検証された、組織が承認したソースからのみアーティファクトのダウンロードを許可し、それ以外はすべてブロックします。

#### #6.4.1 レベル: 1 役割: D/V

モデルの重み、データセット、およびコンテナが承認されたドメインまたは内部レジストリからのみダウンロードされることを確認してください。

#### #6.4.2 レベル: 1 役割: D/V

Sigstore/Cosignの署名がアーティファクトをローカルにキャッシュする前に、公開者の身元を検証することを確認してください。

#### #6.4.3 レベル: 2 役割: D

信頼されたソースポリシーを適用するために、出口プロキシが認証されていないアーティファクトのダウンロードをブロックしていることを確認してください。

**#6.4.4 レベル: 2 役割: V**

リポジトリのホワイトリストが四半期ごとにレビューされており、各エントリーに対するビジネス上の正当性の証拠があることを確認してください。

**#6.4.5 レベル: 3 役割: V**

ポリシー違反がアーティファクトの隔離および依存するパイプラインのロールバックをトリガーすることを確認してください。

## C6.5 サードパーティデータセットリスク評価

外部データセットを毒性、バイアス、法的コンプライアンスの観点から評価し、そのライフサイクルを通じて監視する。

**#6.5.1 レベル: 1 役割: D/V**

外部データセットがポイズニングリスクスコアリング(例:データフィンガープリンティング、外れ値検出)を受けていることを検証してください。

**#6.5.2 レベル: 1 役割: D**

バイアスマトリクス(人口統計的パリティ、平等機会)がデータセット承認前に計算されていることを確認してください。

**#6.5.3 レベル: 2 役割: V**

データセットの出所、系統、およびライセンス条件がML-BOMエントリに記録されていることを確認してください。

**#6.5.4 レベル: 2 役割: V**

定期的なモニタリングがホストされたデータセットのドリフトや破損を検出することを確認する。

**#6.5.5 レベル: 3 役割: D**

トレーニング前に、自動スクラビングによって許可されていないコンテンツ(著作権、個人識別情報)が削除されていることを確認してください。

## C6.6 サプライチェーン攻撃モニタリング

CVEフィード、監査ログ分析、レッドチームシミュレーションを通じて、サプライチェーンの脅威を早期に検出します。

**#6.6.1 レベル: 1 役割: V**

CI/CDの監査ログが、異常なパッケージブルや改ざんされたビルドステップの検出のためにSIEMにストリーミングされていることを検証してください。

**#6.6.2 レベル: 2 役割: D**

インシデント対応プレイブックに、侵害されたモデルやライブラリのロールバック手順が含まれていることを確認してください。

**#6.6.3 レベル: 3 役割: V**

アラートトリアージにおいて、脅威インテリジェンスの強化がML特有の指標(例:モデル汚染のIoC)をタグ付けしていることを検証する。

## C6.7 モデルアーティファクトのための ML-BOM

詳細なML特化型SBOM(ML-BOM)を生成および署名し、下流の利用者が展開時にコンポーネントの整合性を検証できるようにします。

### #6.7.1 レベル: 1 役割: D/V

すべてのモデルアーティファクトが、データセット、重み、ハイパーパラメータ、およびライセンスを一覧表示するML-BOMを公開していることを確認してください。

### #6.7.2 レベル: 1 役割: D/V

ML-BOMの生成とCosign署名がCIで自動化されており、マージに必須であることを確認してください。

### #6.7.3 レベル: 2 役割: D

ML-BOMの完全性チェックが、コンポーネントのメタデータ(ハッシュ、ライセンス)が欠落している場合にビルドを失敗させることを確認してください。

### #6.7.4 レベル: 2 役割: V

ダウンストリームの利用者が API を介して ML-BOM をクエリし、デプロイ時にインポートされたモデルを検証できることを確認してください。

### #6.7.5 レベル: 3 役割: V

ML-BOMがバージョン管理されており、不正な変更を検出するために差分が確認されていることを検証してください。

## 参考文献

- OWASP LLM03:2025 Supply Chain
- MITRE ATLAS : Supply Chain Compromise
- SBOM Overview – CISA
- CycloneDX – Machine Learning Bill of Materials

# C7 モデルの動作、出力制御および安全保証

## 制御目標

モデルの出力は構造化され、信頼性があり、安全で説明可能でなければならず、運用中は継続的に監視される必要があります。これにより、幻覚(ハルシネーション)、プライバシー漏洩、有害コンテンツ、および暴走行動が減少し、ユーザーの信頼と規制遵守が向上します。

### C7.1 出力フォーマットの強制

厳格なスキーマ、制約付きデコーディング、および下流の検証により、不正な形式や悪意のあるコンテンツが拡散する前に阻止します。

#### #7.1.1 レベル: 1 役割: D/V

システムプロンプトにレスポンススキーマ(例:JSONスキーマ)が提供されていることを確認し、すべての出力が自動的に検証されるようにします。スキーマに準拠しない出力は修正または拒否されるようにします。

#### #7.1.2 レベル: 1 役割: D/V

オーバーフローやプロンプトインジェクションのサイドチャネルを防ぐために、制約付きデコーディング(ストップトークン、正規表現、最大トークン数)が有効になっていることを確認してください。

#### #7.1.3 レベル: 2 役割: D/V

下流のコンポーネントが出力を信頼できないものとして扱い、それらをスキーマやインジェクション安全なデシリアライザで検証していることを確認してください。

#### #7.1.4 レベル: 3 役割: V

不適切な出力イベントがログに記録され、レート制限され、監視に表示されることを確認してください。

### C7.2 幻覚検出と緩和

不確実性の推定とフォールバック戦略は、捏造された回答を抑制します。

#### #7.2.1 レベル: 1 役割: D/V

トークンレベルの対数確率、アンサンブル自己一貫性、または微調整された幻覚検出器が各回答に信頼度スコアを割り当てるかを確認してください。

#### #7.2.2 レベル: 1 役割: D/V

設定可能な信頼度閾値を下回る応答がフォールバックワークフロー(例:リトリーバル強化生成、二次モデル、または人間によるレビュー)をトリガーすることを検証します。

#### #7.2.3 レベル: 2 役割: D/V

幻覚インシデントが根本原因のメタデータでタグ付けされ、ポストモーテムおよびファインチューニングパイプラインに供給されているかを確認してください。

#### #7.2.4 レベル: 3 役割: D/V

主要なモデルまたは知識ベースの更新後に、しきい値と検出器が再キャリブレーションされているかを確認

してください。

#### #7.2.5 レベル: 3 役割: V

ダッシュボードのビジュアライゼーションが幻覚率を追跡していることを確認してください。

## C7.3 出力の安全性およびプライバシーフィルタリング

ポリシーフィルターとレッドチームのカバレッジは、ユーザーと機密データを保護します。

#### #7.3.1 レベル: 1 役割: D/V

ポリシーに沿ったヘイト、嫌がらせ、自傷行為、過激派、性的に露骨なコンテンツを、生成前および生成後の分類器がブロックしていることを確認してください。

#### #7.3.2 レベル: 1 役割: D/V

すべての応答でPII/PCI検出と自動編集が実行されることを確認してください。違反が発生した場合はプライバシーアインシデントを報告します。

#### #7.3.3 レベル: 2 役割: D

機密性タグ(例:企業秘密)が複数のモダリティにわたって伝播し、テキスト、画像、コードにおける漏洩を防止することを確認してください。

#### #7.3.4 レベル: 3 役割: D/V

フィルターバイパスの試みや高リスクの分類には、二次承認またはユーザーの再認証が必要であることを確認してください。

#### #7.3.5 レベル: 3 役割: D/V

フィルタリングの閾値が法的管轄区域およびユーザーの年齢・役割のコンテキストを反映していることを確認する。

## C7.4 出力およびアクション制限

レート制限と承認ゲートは、悪用や過度な自律性を防止します。

#### #7.4.1 レベル: 1 役割: D

429エラー発生時に指數関数的なバックオフを使用して、ユーザーごとおよびAPIキーごとのクオータがリクエスト数、トークン数、およびコストを制限していることを検証してください。

#### #7.4.2 レベル: 1 役割: D/V

特権操作(ファイル書き込み、コード実行、ネットワーク呼び出し)がポリシーベースの承認またはヒューマンインザループを必要とするかを確認してください。

#### #7.4.3 レベル: 2 役割: D/V

クロスモーダル整合性チェックが、同じリクエストに対して生成された画像、コード、テキストが悪意のあるコンテンツの密輸に使用されないことを保証することを検証してください。

#### #7.4.4 レベル: 2 役割: D

エージェントの委任の深さ、再帰制限、および許可されたツールリストが明示的に構成されていることを確認してください。

#### #7.4.5 レベル: 3 役割: V

制限違反がSIEM取り込み用の構造化されたセキュリティイベントを発生させることを検証してください。

## C7.5 出力の説明可能性

透明なシグナルは、ユーザーの信頼と内部デバッグを向上させます。

### #7.5.1 レベル: 2 役割: D/V

リスク評価が適切と判断した場合に、ユーザー向けの信頼度スコアまたは簡潔な推論概要が表示されることを確認してください。

### #7.5.2 レベル: 2 役割: D/V

生成された説明が、機密性の高いシステムプロンプトや専有データを明かさないように確認してください。

### #7.5.3 レベル: 3 役割: D

システムがトークンレベルのログ確率またはアテンションマップをキャプチャし、許可された検査のために保存していることを確認してください。

### #7.5.4 レベル: 3 役割: V

説明可能性の成果物が監査可能性のためにモデルのリリースとともにバージョン管理されていることを確認してください。

## C7.6 監視統合

リアルタイムオブザーバビリティは、開発と本番環境の間のループを閉じます。

### #7.6.1 レベル: 1 役割: D

メトリクス(スキーマ違反、幻覚率、トキシシティ、PII漏洩、レイテンシ、コスト)が中央監視プラットフォームにストリームされることを検証してください。

### #7.6.2 レベル: 1 役割: V

各安全指標に対してアラートしきい値が定義されており、オンコールのエスカレーション経路が確立されていることを確認してください。

### #7.6.3 レベル: 2 役割: V

ダッシュボードが出力異常をモデル／バージョン、機能フラグ、および上流データの変更と関連付けていることを確認してください。

### #7.6.4 レベル: 2 役割: D/V

監視データが文書化されたMLOpsワークフロー内で再トレーニング、ファインチューニング、またはルールの更新にフィードバックされることを確認してください。

### #7.6.5 レベル: 3 役割: V

監視パイプラインがペネトレーションテストを受けており、機密ログの漏洩を防ぐためにアクセス制御されていることを確認してください。

## 7.7 生成メディアの安全対策

AIシステムが法律に違反する、有害な、または許可されていないメディアコンテンツを生成しないように、ポリシー制約の強制、出力の検証、および追跡可能性を確保する。

**#7.7.1 レベル: 1 役割: D/V**

システムプロンプトおよびユーザー指示が、違法、有害、または非合意のディープフェイクメディア（例：画像、動画、音声）の生成を明示的に禁止していることを確認してください。

**#7.7.2 レベル: 2 役割: D/V**

プロンプトがなりすましの生成、性的に露骨なディープフェイク、または同意なしに実在の個人を描写するメディアの生成を試みていないかどうかを検証してください。

**#7.7.3 レベル: 2 役割: V**

システムが著作権で保護されたメディアの不正複製を防止するために、知覚ハッシュ、ウォーターマーク検出、またはフィンガープリントングを使用していることを確認してください。

**#7.7.4 レベル: 3 役割: D/V**

すべての生成されたメディアが暗号的に署名されていること、透かしが入っていること、または改ざん防止の起源情報メタデータが埋め込まれており、下流でのトレーサビリティが確保されていることを検証してください。

**#7.7.5 レベル: 3 役割: V**

バイパス試行（例：プロンプトの難読化、スラング、敵対的表現）が検出され、ログに記録され、レート制限されていることを検証する。繰り返される悪用は監視システムに報告される。

## 参考文献

- NIST AI Risk Management Framework
- ISO/IEC 42001:2023 – AI Management System
- OWASP Top-10 for Large Language Model Applications (2025)
- Practical Techniques to Constrain LLM Output
- Dataiku – Structured Text Generation Guide
- VL-Uncertainty: Detecting Hallucinations
- HaDeMiF: Hallucination Detection & Mitigation
- Building Confidence in LLM Outputs
- Explainable AI & LLMs
- Sensitive Information Disclosure in LLMs
- OpenAI Rate-Limit & Exponential Back-off
- Arize AI – LLM Observability Platform

# C8 メモリ、埋め込み、ベクターデータベースのセキュリティ

## 制御目標

埋め込みとベクターストアは、現代のAIシステムの「ライブメモリ」として機能し、ユーザーから提供されたデータを継続的に受け入れ、検索拡張生成(Retrieval-Augmented Generation, RAG)を通じてモデルのコンテキストに再提示します。管理されていない場合、このメモリは個人識別情報(PII)を漏洩したり、同意に違反したり、元のテキストを再構成するために逆解析される可能性があります。この制御群の目的は、アクセス権を最小特権に制限し、埋め込みがプライバシー保護され、保存されたベクトルが期限切れになるか要求に応じて取り消し可能であり、ユーザーごとのメモリが他のユーザーのプロンプトや完了結果に混入しないように、メモリパイプラインとベクターデータベースを強化することです。

### C8.1 メモリおよびRAGインデックスへのアクセス制御

すべてのベクターコレクションに対して細粒度のアクセス制御を適用してください。

#### #8.1.1 レベル: 1 役割: D/V

テナント、コレクション、またはドキュメントタグごとに、行/名前空間レベルのアクセス制御ルールが挿入、削除、およびクエリ操作を制限していることを確認してください。

#### #8.1.2 レベル: 1 役割: D/V

APIキーやJWTにスコープ付きクレーム(例:コレクションID、操作動詞)が含まれていることを確認し、少なくとも四半期ごとにローテーションされていることを検証してください。

#### #8.1.3 レベル: 2 役割: D/V

特権昇格の試み(例:クロスネームスペースの類似性クエリ)が検出され、5分以内にSIEMにログが記録されることを検証してください。

#### #8.1.4 レベル: 2 役割: D/V

ベクターデータベースが監査ログにサブジェクト識別子、操作、ベクターID/ネームスペース、類似度閾値、および結果数を記録していることを確認してください。

#### #8.1.5 レベル: 3 役割: V

エンジンがアップグレードされるか、インデックスシャーディングルールが変更されるたびに、アクセス決定がバイパスの欠陥についてテストされていることを検証してください。

### C8.2 埋め込みのサニタイズおよび検証

ベクトル化する前に、テキストを事前にPII(個人識別情報)についてスクリーニングし、編集または仮名化し、必要に応じて埋め込み後処理で残留信号を除去します。

#### #8.2.1 レベル: 1 役割: D/V

PIIおよび規制対象データが自動分類器によって検出され、埋め込み前にマスク、トークン化、または削除されることを検証してください。

#### #8.2.2 レベル: 1 役割: D

埋め込みパイプラインが、インデックスを汚染する可能性のある実行可能コードや非UTF-8のアーティファクトを含む入力を拒否または隔離することを検証してください。

#### #8.2.3 レベル: 2 役割: D/V

既知のPIIトークンからの距離が設定可能な閾値を下回る文の埋め込みに対して、ローカルまたはメトリック差分プライバシーのサンタイズ処理が適用されていることを検証してください。

#### #8.2.4 レベル: 2 役割: V

サンタイズの有効性(例: PII編集のリコール、意味のずれ)が、ベンチマークコーパスに対して少なくとも年に2回検証されていることを確認してください。

#### #8.2.5 レベル: 3 役割: D/V

サンタイズ設定がバージョン管理されており、変更がピアレビューを受けることを確認してください。

## C8.3 メモリの有効期限、取り消し、および削除

GDPRの「忘れられる権利」や類似の法律は、適時の消去を要求するため、ベクトルストアはTTL、ハードディレート、およびトーンストーニングをサポートし、取り消されたベクトルが回復または再インデックスされないようにする必要があります。

#### #8.3.1 レベル: 1 役割: D/V

すべてのベクトルおよびメタデータレコードが、TTL(Time To Live)または自動クリーンアップジョブによって尊重される明示的な保持ラベルを持っていることを確認してください。

#### #8.3.2 レベル: 1 役割: D/V

ユーザーによる削除リクエストが、30日以内にベクター、メタデータ、キャッシュコピー、および派生インデックスを完全に削除することを確認してください。

#### #8.3.3 レベル: 2 役割: D

ハードウェアが対応している場合は論理的削除に続いてストレージブロックの暗号的シミュレーティングが行われること、またはキーボルトの鍵破棄によって行われることを検証してください。

#### #8.3.4 レベル: 3 役割: D/V

有効期限切れのベクトルが期限切れから500ミリ秒以内に最近傍検索結果から除外されることを検証してください。

## C8.4 埋め込み反転および漏洩の防止

最近の防御策—ノイズ重ね合わせ、射影ネットワーク、プライバシーニューロン振動、およびアプリケーション層の暗号化—は、トークンレベルの逆転率を5%以下に抑えることができます。

#### #8.4.1 レベル: 1 役割: V

逆転攻撃、メンバーシップ攻撃、および属性推測攻撃を網羅した正式な脅威モデルが存在し、毎年レビューされていることを確認してください。

#### #8.4.2 レベル: 2 役割: D/V

アプリケーション層の暗号化または検索可能な暗号化が、インフラ管理者やクラウドスタッフによるベクトルの直接読み取りから保護していることを確認してください。

#### #8.4.3 レベル: 3 役割: V

防御パラメータ(DPの $\epsilon$ 、ノイズの $\sigma$ 、射影ランクk)がプライバシー $\geq 99\%$ のトーケン保護とユーティリティ $\leq 3\%$ の精度低下のバランスを保っていることを検証してください。

#### #8.4.4 レベル: 3 役割: D/V

モデル更新のリリースゲートの一部として、反転耐性メトリクスが含まれていることを確認し、回帰予算が定義されていることを確認してください。

## C8.5 ユーザー固有メモリのスコープ強制

クロステナントの情報漏洩は依然として主要なRAGリスクであり、不適切にフィルタリングされた類似性クエリが他の顧客のプライベートドキュメントを露出させる可能性があります。

#### #8.5.1 レベル: 1 役割: D/V

すべての検索クエリがLLMプロンプトに渡される前に、テナント／ユーザーIDによってポストフィルタリングされていることを確認してください。

#### #8.5.2 レベル: 1 役割: D

コレクション名または名前空間付きIDがユーザーまたはテナントごとにソルト処理されていることを確認し、スコープ間でベクトルが衝突しないようにしてください。

#### #8.5.3 レベル: 2 役割: D/V

類似度が設定可能な距離閾値を超えるか、かつ呼び出し元のスコープ外である結果は破棄され、セキュリティアラートが発生することを検証してください。

#### #8.5.4 レベル: 2 役割: V

マルチテナントストレステストが、スコープ外のドキュメントを取得しようとする敵対的なクエリをシミュレートし、情報漏洩が全くないことを実証していることを確認してください。

#### #8.5.5 レベル: 3 役割: D/V

暗号鍵がテナントごとに分離されていることを確認し、物理ストレージが共有されている場合でも暗号的な分離が確保されていることを保証してください。

## C8.6 高度なメモリシステムセキュリティ

エピソード記憶、意味記憶、作業記憶を含む高度なメモリアーキテクチャに対する、特定の分離および検証要件を備えたセキュリティコントロール。

#### #8.6.1 レベル: 1 役割: D/V

異なるメモリタイプ(エピソード記憶、意味記憶、作業記憶)が、ロールベースアクセス制御、個別の暗号化キー、および各メモリタイプのアクセスパターンの文書化により、分離されたセキュリティコンテキストを持っていることを検証してください。

#### #8.6.2 レベル: 2 役割: D/V

記憶統合プロセスが、コンテンツのサニタイズ、ソースの検証、および格納前の整合性チェックを通じて悪意のある記憶の注入を防止するためのセキュリティ検証を含んでいることを確認してください。

**#8.6.3 レベル: 2 役割: D/V**

メモリ検索クエリが検証およびサニタイズされ、不正な情報抽出を防ぐために、クエリパターンの分析、アクセス制御の実施、および結果のフィルタリングが行われていることを確認してください。

**#8.6.4 レベル: 3 役割: D/V**

キーの削除、複数回上書き、または検証証明書付きのハードウェアベースの安全な削除を使用して、暗号消去の保証がある機密情報を安全に削除するメモリ忘却メカニズムを検証してください。

**#8.6.5 レベル: 3 役割: D/V**

メモリシステムの整合性が、チェックサム、監査ログ、およびメモリ内容が通常の操作範囲外で変更された際の自動アラートを通じて、不正な変更や破損が継続的に監視されていることを検証する。

## 参考文献

- [Vector database security: Pinecone – IronCore Labs](#)
- [Securing the Backbone of AI: Safeguarding Vector Databases and Embeddings – Privacera](#)
- [Enhancing Data Security with RBAC of Qdrant Vector Database – AI Advances](#)
- [Mitigating Privacy Risks in LLM Embeddings from Embedding Inversion – arXiv](#)
- [DPPN: Detecting and Perturbing Privacy-Sensitive Neurons – OpenReview](#)
- [Art. 17 GDPR – Right to Erasure](#)
- [Sensitive Data in Text Embeddings Is Recoverable – Tonic.ai](#)
- [PII Identification and Removal – NVIDIA NeMo Docs](#)
- [De-identifying Sensitive Data – Google Cloud DLP](#)
- [Remove PII from Conversations Using Sensitive Information Filters – AWS Bedrock Guardrails](#)
- [Think Your RAG Is Secure? Think Again – Medium](#)
- [Design a Secure Multitenant RAG Inferencing Solution – Microsoft Learn](#)
- [Best Practices for Multi-Tenancy RAG with Milvus – Milvus Blog](#)

# 9 自律的オーケストレーションとエージェント行動のセキュリティ

## 制御目標

自律型またはマルチエージェントAIシステムが、明確に意図され、認証され、監査可能であり、かつ制限されたコストおよびリスクの閾値内にあるアクションのみを実行できるようにします。これにより、自律システムの侵害、ツールの誤使用、エージェントループの検出、通信のハイジャック、アイデンティティのなりすまし、群れの操作、および意図の操作などの脅威から保護されます。

### 9.1 エージェントのタスク計画と再帰予算

再帰的なプランを制限し、特権アクションには人間のチェックポイントを強制する。

#### #9.1.1 レベル: 1 役割: D/V

最大再帰深度、幅、実時間、トークン、および各エージェント実行あたりの金銭的コストが中央で設定され、バージョン管理されていることを確認してください。

#### #9.1.2 レベル: 1 役割: D/V

特権的または不可逆的な操作(例:コードコミット、財務の振替)が実行される前に、監査可能なチャネルを通じて明確な人間の承認を必要とすることを確認してください。

#### #9.1.3 レベル: 2 役割: D

リアルタイムリソースモニターが任意の予算閾値を超えた場合にサーキットブレーカーの割り込みをトリガーし、さらなるタスクの拡張を停止することを検証してください。

#### #9.1.4 レベル: 2 役割: D/V

サーキットブレーカーイベントがエージェントID、トリガー条件、およびフォレンジックレビューのためにキャプチャされたプラン状態とともにログに記録されていることを確認してください。

#### #9.1.5 レベル: 3 役割: V

セキュリティテストが予算枯渀および無制御プランのシナリオをカバーし、データ損失なしに安全に停止できることを確認してください。

#### #9.1.6 レベル: 3 役割: D

予算ポリシーがポリシー・アズ・コードとして表現され、構成のドリフトを防ぐためにCI/CDで強制されていることを確認してください。

### 9.2 ツールプラグインのサンドボックス化

不正なシステムアクセスやコード実行を防ぐために、ツールの相互作用を分離してください。

#### #9.2.1 レベル: 1 役割: D/V

すべてのツール／プラグインが、最小権限のファイルシステム、ネットワーク、およびシステムコールポリシーを備えたOS、コンテナ、またはWASMレベルのサンドボックス内で実行されていることを確認してください。

#### #9.2.2 レベル: 1 役割: D/V

サンドボックスのリソース割り当て上限(CPU、メモリ、ディスク、ネットワーク送信)および実行タイムアウトが適切に適用され、ログに記録されていることを確認してください。

#### #9.2.3 レベル: 2 役割: D/V

ツールのバイナリまたは記述子がデジタル署名されていることを確認し、ロード前に署名が検証されることを保証してください。

#### #9.2.4 レベル: 2 役割: V

サンドボックスのテレメトリがSIEMにストリーミングされていることを確認する。異常(例: 試みられたアウトバウンド接続)がアラートを発生させる。

#### #9.2.5 レベル: 3 役割: V

高リスクのプラグインが本番環境に展開される前に、セキュリティレビューおよびペネトレーションテストを受けていることを確認してください。

#### #9.2.6 レベル: 3 役割: D/V

サンドボックス脱出の試みが自動的にブロックされ、問題のあるプラグインが調査を待つ間、隔離されることを確認してください。

## 9.3 自律ループとコスト制限

エージェント間の制御不能な再帰およびコスト爆発を検出して停止します。

#### #9.3.1 レベル: 1 役割: D/V

インターフェースエージェントコールが、ランタイムが減算および強制するホップリミットまたはTTLを含んでいることを確認してください。

#### #9.3.2 レベル: 2 役割: D

エージェントが自己呼び出しや循環パターンを検出するために、一意の呼び出しグラフIDを維持していることを確認してください。

#### #9.3.3 レベル: 2 役割: D/V

累積コンピュートユニットおよび支出カウンターがリクエストチェーンごとに追跡されていることを検証する; 制限を超えるとチェーンが中止される。

#### #9.3.4 レベル: 3 役割: V

正式解析またはモデル検査によって、エージェントプロトコルにおける無限再帰が存在しないことを検証する。

#### #9.3.5 レベル: 3 役割: D

ループ中断イベントがアラートを生成し、継続的改善の指標にフィードバックされることを検証してください。

## 9.4 プロトコルレベルの誤用防止

エージェントと外部システム間の通信チャネルを安全にし、乗っ取りや操作を防止する。

#### #9.4.1 レベル: 1 役割: D/V

すべてのエージェント間およびエージェントからツールへのメッセージが認証されていること(例: 相互TLSまたはJWT)およびエンドツーエンドで暗号化されていることを確認してください。

#### #9.4.2 レベル: 1 役割: D

スキーマが厳密に検証されていることを確認し、未知のフィールドや不正な形式のメッセージが拒否されることを保証してください。

**#9.4.3 レベル: 2 役割: D/V**

整合性チェック(MACやデジタル署名)がツールパラメータを含むメッセージペイロード全体をカバーしていることを確認してください。

**#9.4.4 レベル: 2 役割: D**

リプレイ保護(ノンスまたはタイムスタンプウインドウ)がプロトコル層で強制されていることを確認してください。

**#9.4.5 レベル: 3 役割: V**

プロトコル実装がインジェクションやデシリアライズの脆弱性に対してファジングおよび静的解析を受けていることを確認してください。

## 9.5 エージェントの識別と改ざん防止

操作が誰によるものか特定可能であり、変更が検出可能であることを保証してください。

**#9.5.1 レベル: 1 役割: D/V**

各エージェントインスタンスが固有の暗号学的識別子(鍵ペアまたはハードウェアルート認証情報)を保持していることを検証します。

**#9.5.2 レベル: 2 役割: D/V**

すべてのエージェントのアクションが署名およびタイムスタンプ付きであることを検証し、ログには否認防止のための署名が含まれていることを確認してください。

**#9.5.3 レベル: 2 役割: V**

改ざん検知可能なログが追記専用または一度書き込み専用の媒体に保存されていることを確認してください。

**#9.5.4 レベル: 3 役割: D**

識別キーが定義されたスケジュールおよび侵害の指標に基づいてローテーションされることを検証する。

**#9.5.5 レベル: 3 役割: D/V**

なりすましやキー衝突の試みが発生した場合、影響を受けたエージェントが即座に隔離されることを確認してください。

## 9.6 マルチエージェントスウォームリスク削減

集団行動に伴う危険を、分離と形式的な安全モデリングによって軽減する。

**#9.6.1 レベル: 1 役割: D/V**

異なるセキュリティドメインで動作するエージェントが、分離されたランタイムサンドボックスまたはネットワークセグメントで実行されていることを確認してください。

**#9.6.2 レベル: 3 役割: V**

スウォーム行動が展開前にライセンスと安全性のためにモデル化され、形式的に検証されていることを確認してください。

**#9.6.3 レベル: 3 役割: D**

ランタイムモニターが新たに発生する安全でないパターン(例えば、振動、デッドロック)を検出し、是正措置を開始することを確認する。

## 9.7 ユーザーおよびツールの認証／認可

すべてのエージェント起動アクションに対して堅牢なアクセス制御を実装してください。

### #9.7.1 レベル: 1 役割: D/V

エージェントがダウンストリームシステムに対して第一級プリンシパルとして認証し、エンドユーザーの資格情報を決して再利用しないことを確認してください。

### #9.7.2 レベル: 2 役割: D

細かい粒度の認可ポリシーが、エージェントが呼び出せるツールおよび提供できるパラメータを制限していることを検証してください。

### #9.7.3 レベル: 2 役割: V

権限チェックがセッション開始時だけでなく、すべての呼び出し時に再評価されること（継続的な認可）を確認してください。

### #9.7.4 レベル: 3 役割: D

委任された権限が自動的に期限切れとなり、タイムアウトまたはスコープの変更後に再同意が必要であることを確認してください。

## 9.8 エージェント間通信のセキュリティ

盗聴や改ざんを防ぐために、すべてのエージェント間メッセージを暗号化し、完全性を保護してください。

### #9.8.1 レベル: 1 役割: D/V

エージェントチャネルに対して、相互認証および完全前方秘匿性暗号化(例:TLS 1.3)が必須であることを検証してください。

### #9.8.2 レベル: 1 役割: D

メッセージの整合性と発信元が処理前に検証されていることを確認してください。検証に失敗した場合は警告を発し、メッセージを破棄します。

### #9.8.3 レベル: 2 役割: D/V

通信メタデータ(タイムスタンプ、シーケンス番号)がフォレンジック再構築をサポートするために記録されていることを確認してください。

### #9.8.4 レベル: 3 役割: V

プロトコル状態機械が安全でない状態に陥らないことを、形式検証またはモデル検査によって確認してください。

## 9.9 意図検証と制約強制

エージェントの行動がユーザーの明示された意図およびシステムの制約と一致していることを検証する。

### #9.9.1 レベル: 1 役割: D

事前実行制約ソルバーが提案された行動をハードコーディングされた安全性およびポリシールールと照合していることを確認してください。

#### #9.9.2 レベル: 2 役割: D/V

高影響の操作(財務的、破壊的、プライバシーに敏感なもの)については、開始ユーザーからの明示的な意図の確認を必要とすることを検証してください。

#### #9.9.3 レベル: 2 役割: V

事後条件チェックが完了したアクションが意図した効果を副作用なしに達成していることを検証し、不一致があればロールバックを引き起こすことを確認してください。

#### #9.9.4 レベル: 3 役割: V

形式的手法(例:モデル検査、定理証明)またはプロパティベースのテストによって、エージェントの計画がすべての宣言された制約を満たしていることを検証します。

#### #9.9.5 レベル: 3 役割: D

意図の不一致や制約違反の事例が継続的な改善サイクルおよび脅威インテリジェンスの共有に反映されることを確認してください。

## 9.10 エージェント推論戦略のセキュリティ

ReAct、Chain-of-Thought、および Tree-of-Thoughts アプローチを含むさまざまな推論戦略の安全な選択と実行。

#### #9.10.1 レベル: 1 役割: D/V

推論戦略の選択が決定論的な基準(入力の複雑さ、タスクの種類、セキュリティコンテキスト)を使用していることを検証し、同一のセキュリティコンテキスト内で同じ入力が同一の戦略選択を生成することを確認してください。

#### #9.10.2 レベル: 1 役割: D/V

各推論戦略(ReAct、Chain-of-Thought、Tree-of-Thoughts)に対して、それぞれの認知アプローチに特化した入力検証、出力サニタイズ、および実行時間制限が設けられていることを確認してください。

#### #9.10.3 レベル: 2 役割: D/V

推論戦略の遷移が、入力特性、選択基準の値、および実行メタデータを含む完全なコンテキストとともに記録されていることを検証し、監査証跡の再構築が可能であることを確認してください。

#### #9.10.4 レベル: 2 役割: D/V

Tree-of-Thoughts推論には、ポリシー違反、リソース制限、安全境界が検出された場合に探索を終了する分岐剪定メカニズムが含まれていることを確認してください。

#### #9.10.5 レベル: 2 役割: D/V

ReAct(Reason-Act-Observe)サイクルには、各フェーズでの検証チェックポイントが含まれていることを確認してください: 推論ステップの検証、行動の承認、次の段階に進む前の観察のサニタイズ。

#### #9.10.6 レベル: 3 役割: D/V

推論戦略のパフォーマンス指標(実行時間、リソース使用量、出力品質)が、設定されたしきい値を超えて逸脱した場合に自動アラートで監視されていることを確認してください。

#### #9.10.7 レベル: 3 役割: D/V

複数の戦略を組み合わせたハイブリッド推論アプローチが、いずれの構成戦略の入力検証および出力制約を維持し、いかなるセキュリティ制御も回避しないことを確認してください。

#### #9.10.8 レベル: 3 役割: D/V

推論戦略のセキュリティテストには、不正な入力によるファジング、戦略の切り替えを強制するように設計された敵対的プロンプト、および各認知アプローチに対する境界条件テストが含まれていることを検証してください。

## 9.11 エージェントライフサイクル状態管理とセキュリティ

暗号化された監査証跡と定義された復旧手順を用いた安全なエージェントの初期化、状態遷移、および終了。

### #9.11.1 レベル: 1 役割: D/V

エージェントの初期化が、ハードウェアに基づく認証情報による暗号学的アイデンティティの確立と、エージェントID、タイムスタンプ、構成ハッシュ、および初期化パラメータを含む不変の起動監査ログを含んでいることを検証してください。

### #9.11.2 レベル: 2 役割: D/V

エージェントの状態遷移が暗号的に署名され、タイムスタンプが付与されており、トリガーイベント、前の状態ハッシュ、新しい状態ハッシュ、および実施されたセキュリティ検証を含む完全なコンテキストとともにログ記録されていることを検証してください。

### #9.11.3 レベル: 2 役割: D/V

エージェントのシャットダウン手順に、暗号消去または多重上書きを用いた安全なメモリ消去、証明書機関への通知を伴う資格情報の取り消し、および改ざん検知可能な終了証明書の生成が含まれていることを確認してください。

### #9.11.4 レベル: 3 役割: D/V

エージェントの回復メカニズムが暗号学的チェックサム(最低でもSHA-256)を使用して状態の整合性を検証し、破損が検出された場合には既知の正常な状態にロールバックし、自動通知および手動承認の要件を備えていることを確認してください。

### #9.11.5 レベル: 3 役割: D/V

エージェントの持続性メカニズムが、エージェントごとのAES-256キーで機密状態データを暗号化し、設定可能なスケジュール(最大90日)で安全なキー回転をゼロダウンタイムで展開することを検証してください。

## 9.12 ツール統合セキュリティフレームワーク

定義されたリスク評価および承認プロセスを伴う、動的ツールのロード、実行、および結果検証に関するセキュリティ制御。

### #9.12.1 レベル: 1 役割: D/V

ツール記述子に、必要な権限(読み取り/書き込み/実行)、リスクレベル(低/中/高)、リソース制限(CPU、メモリ、ネットワーク)、およびツールマニフェストに記載された検証要件を指定するセキュリティメタデータが含まれていることを確認してください。

### #9.12.2 レベル: 1 役割: D/V

ツールの実行結果が期待されるスキーマ(JSONスキーマ、XMLスキーマ)およびセキュリティポリシー(出力のサニタイズ、データ分類)に対して検証されていることを確認し、タイムアウト制限およびエラー処理手順と統合する前に検証を行います。

### #9.12.3 レベル: 2 役割: D/V

ツールのインターラクションログに、権限使用状況、データアクセスパターン、実行時間、リソース消費、およびリターンコードを含む詳細なセキュリティコンテキストが記録されていることを検証し、SIEM統合のための構造化ログを確保してください。

**#9.12.4 レベル: 2 役割: D/V**

動的ツール読み込み機構がPKIインフラストラクチャを使用してデジタル署名を検証し、実行前にサンドボックス分離および権限検証を伴う安全な読み込みプロトコルを実装していることを確認してください。

**#9.12.5 レベル: 3 役割: D/V**

新しいバージョンに対してツールのセキュリティ評価が自動的に開始されることを確認し、必須の承認ゲートには静的解析、動的テスト、およびセキュリティチームのレビューが含まれ、承認基準とSLA要件が文書化されていることを確認してください。

## C9.13 モデルコンテキストプロトコル (MCP) セキュリティ

文脈の混乱、不正なツール呼び出し、またはテナント間でのデータ露出を防ぐために、MCPベースのツールおよびリソース統合の安全な検出、認証、認可、通信、および使用を確実に行います。

### コンポーネントの整合性とサプライチェーンの衛生管理

**#9.13.1 レベル: 1 役割: D/V**

MCPサーバー、クライアント、およびツールの実装が、不安全な関数の露出、安全でないデフォルト設定、認証の欠如、または入力検証の欠如を特定するために、手動でレビューされるか自動的に分析されていることを確認してください。

**#9.13.2 レベル: 1 役割: D/V**

外部またはオープンソースのMCPサーバーやパッケージが統合前に自動化された脆弱性およびサプライチェーンスキャン(例:SCA)を受けていること、また既知の重大な脆弱性を持つコンポーネントが使用されていないことを確認してください。

**#9.13.3 レベル: 1 役割: D/V**

MCPサーバーおよびクライアントコンポーネントが信頼できるソースからのみ取得されていることを確認し、署名、チェックサム、または安全なパッケージメタデータを使用して検証し、改ざんされたビルドや署名されていないビルドを拒否してください。

### 認証と認可

**#9.13.4 レベル: 2 役割: D/V**

MCPクライアントとサーバーが強力な非ユーザー認証情報(例:mTLS、署名付きトークン、プラットフォーム発行の識別子)を用いて相互認証を行い、認証されていないMCPエンドポイントが拒否されることを検証してください。

**#9.13.5 レベル: 2 役割: D/V**

MCPサーバーが、所有者、環境、およびリソースの明確な定義を必要とする管理された技術的オンボーディングメカニズムを通じて登録されていることを検証してください。登録されていない、または検出不可能なサーバーは、本番環境で呼び出すことができません。

**#9.13.6 レベル: 2 役割: D/V**

各MCPツールまたはリソースが明示的な認可スコープ(例:読み取り専用、制限されたクエリ、副作用レベル)を定義していることを確認し、エージェントが割り当てられたスコープ外のMCP機能を呼び出せないようにしてください。

### 安全なトランSPORTとネットワーク境界防御

**#9.13.7 レベル: 2 役割: D/V**

認証済みかつ暗号化されたストリーム可能なHTTPが本番環境における主要なMCPトランSPORTとして使用されていることを確認してください。代替トランSPORT(stdio、SSE)は、ローカルまたは厳密に管理された環境に限定され、明確な理由がある場合にのみ使用されます。

**#9.13.8 レベル: 2 役割: D/V**

streamable-HTTP MCPトランSPORTが、認証済みかつ暗号化されたチャネル(TLS 1.3以降)を使用し、証明書検証と前方秘匿性を備えて、ストリームされるMCPメッセージの機密性と完全性を確保していることを検証してください。

**#9.13.9 レベル: 2 役割: D/V**

SSEベースのMCPトランSPORTが、プライベートで認証された内部チャネル内でのみ使用されていることを確認し、TLS、認証、スキーマ検証、ペイロードサイズ制限、およびレート制限を適用してください。SSEエンドポイントはパブリックインターネットに公開してはなりません。

**#9.13.10 レベル: 2 役割: D/V**

MCPサーバーが検証することを確認してください。`Origin` と `Host` DNSリバインディング攻撃を防ぐために、すべてのHTTPベースのトランSPORT(SSEおよびストリーム可能なHTTPを含む)でヘッダーを設定し、信頼できない、ミスマッチした、またはオリジンが欠落しているリクエストを拒否します。

## スキーマ、メッセージ、および入力検証

**#9.13.11 レベル: 2 役割: D/V**

MCPツールおよびリソーススキーマ(例:JSONスキーマや機能記述子)が、スキーマの改ざんや悪意のあるパラメータ変更を防止するために、署名、チェックサム、またはサーバー認証を用いて真正性と完全性が検証されていることを確認してください。

**#9.13.12 レベル: 2 役割: D/V**

すべてのMCPトランSPORTが、メッセージフレーミングの整合性、厳格なスキーマ検証、最大ペイロードサイズの確保、および誤形成、切り詰められた、またはインターリーブされたフレームの拒否を強制し、同期ずれやインジェクション攻撃を防止していることを検証してください。

**#9.13.13 レベル: 2 役割: D/V**

MCPサーバーがすべての関数呼び出しに対して、型チェック、境界チェック、列挙型の強制、および認識されないまたは過大なパラメータの拒否を含む厳格な入力検証を行っていることを確認してください。

## アウトバウンドアクセスとエージェント実行の安全性

**#9.13.14 レベル: 2 役割: D/V**

MCPサーバーは、最小権限の出口ポリシーに従って承認された内部または外部の宛先にのみアウトバウンドリクエストを開始でき、任意のネットワークターゲットや内部クラウドメタデータサービスにはアクセスできないことを検証してください。

**#9.13.15 レベル: 2 役割: D/V**

アウトバウンドMCPアクションが、無制限のエージェント駆動ツール呼び出しや連鎖的な副作用を防ぐために、実行制限(タイムアウト、再帰制限、同時実行制限、サーキットブレーカー)を実装していることを検証してください。

**#9.13.16 レベル: 2 役割: D/V**

MCPのリクエストおよびレスポンスのメタデータ(サーバーID、リソース名、ツール名、セッション識別子、テナント、環境)が完全性保護付きでログに記録されており、法医学的分析のためにエージェントの活動と関連付けられていることを確認してください。

## 輸送制限と高リスク境界管理

**#9.13.17 レベル: 3 役割: D/V**

stdioベースのMCPトランSPORTは、シェルの実行、端末の注入、およびプロセスの生成機能から隔離された、同一場所の单一プロセス開発シナリオに限定されていることを確認してください; stdioは決してネットワークまたはマルチテナントの境界を越えてはなりません。

**#9.13.18 レベル: 3 役割: D/V**

MCPサーバーが許可リストに載っている関数とリソースのみを公開し、ユーザーやモデルによって提供された入力に影響される関数名の動的ディスパッチ、リフレクティブな呼び出し、または実行を禁止していることを検証してください。

**#9.13.19 レベル: 3 役割: D/V**

MCPレイヤーでテナント境界、環境境界(開発/テスト/本番)、およびデータドメイン境界が適用され、テナント間または環境間でのサーバーやリソースの検出が防止されていることを確認してください。

## 参考文献

- MITRE ATLAS tactics ML09
- Circuit-breaker research for AI agents — Zou et al., 2024
- Trend Micro analysis of sandbox escapes in AI agents — Park, 2025
- Auth0 guidance on human-in-the-loop authorization for agents — Martinez, 2025
- Medium deep-dive on MCP & A2A protocol hijacking — ForAISeC, 2025
- Rapid7 fundamentals on spoofing attack prevention — 2024
- Imperial College verification of swarm systems — Lomuscio et al.
- NIST AI Risk Management Framework 1.0, 2023
- WIRED security briefing on encryption best practices, 2024
- OWASP Top 10 for LLM Applications, 2025
- Comprehensive Vulnerability Analysis is Necessary for Trustworthy LLM-MAS
- [How Is LLM Reasoning Distracted by Irrelevant Context? An Analysis Using a Controlled Benchmark] (<https://www.arxiv.org/pdf/2505.18761>)
- Large Language Model Sentinel: LLM Agent for Adversarial Purification
- Securing Agentic AI: A Comprehensive Threat Model and Mitigation Framework for Generative AI Agents
- Model Context Protocol Specification
- Model Context Protocol Tools & Resources Specification
- Model Context Protocol Transport Documentation
- OWASP GenAI Security Project — “A Practical Guide for Securely Using Third-Party MCP Servers 1.0”
- Cloud Security Alliance – Model Context Protocol Security Working Group
- CSA MCP Security: Top 10 Risks
- CSA MCP Security: TTPs & Hardening Guidance

# 10 敵対的ロバストネスとプライバシー防御

## 制御目標

AIモデルが回避、推論、抽出、または毒物攻撃に直面した場合でも、信頼性があり、プライバシーを保護し、悪用に耐性があることを確保してください。

### 10.1 モデルの整合性と安全性

有害またはポリシー違反の出力を防止する。

#### #10.1.1 レベル: 1 役割: D/V

アライメントテストスイート(レッドチームプロンプト、ジャイルブレイクプローブ、不許可コンテンツ)がバージョン管理されており、すべてのモデルリリースで実行されていることを確認してください。

#### #10.1.2 レベル: 1 役割: D

拒否および安全完了ガードレールが適用されていることを確認してください。

#### #10.1.3 レベル: 2 役割: D/V

自動評価者が有害コンテンツ率を測定し、設定された閾値を超える後退を検出することを確認してください。

#### #10.1.4 レベル: 2 役割: D

カウンタージェイルブレイクトレーニングが文書化され、再現可能であることを確認してください。

#### #10.1.5 レベル: 3 役割: V

正式なポリシー遵守証明や認証された監視が重要なドメインをカバーしていることを検証してください。

### 10.2 敵対的事例の耐性強化

操作された入力に対する耐性を高める。堅牢な敵対的訓練とベンチマークスコアリングが現在の最善の方法である。

#### #10.2.1 レベル: 1 役割: D

プロジェクトリポジトリに、再現可能なシードを使用した敵対的トレーニング構成が含まれていることを確認してください。

#### #10.2.2 レベル: 2 役割: D/V

敵対的サンプル検出が本番パイプラインでブロックキングアラートを発生させることを検証してください。

#### #10.2.4 レベル: 3 役割: V

認証済みの堅牢性証明または区間境界証明が、少なくとも最も重要な主要クラスをカバーしていることを検証してください。

#### #10.2.5 レベル: 3 役割: V

回帰テストが適応的攻撃を使用して測定可能な堅牢性の低下がないことを確認していることを検証してください。

## 10.3 メンバーシップ推論緩和

記録がトレーニングデータに含まれていたかどうかを決定する能力を制限する。差分プライバシーと信頼度スコアのマスキングは、現在知られている最も効果的な防御策である。

### #10.3.1 レベル: 1 役割: D

クエリごとのエントロピー正則化または温度スケーリングが過信的な予測を減少させることを検証してください。

### #10.3.2 レベル: 2 役割: D

機微なデータセットに対して、トレーニングが  $\epsilon$ -有界差分プライバシー最適化を適用していることを検証してください。

### #10.3.3 レベル: 2 役割: V

攻撃シミュレーション(シャドウモデルまたはブラックボックス)が、保留データに対して攻撃の  $AUC \leq 0.60$  を示すことを確認してください。

## 10.4 モデル反転耐性

プライベート属性の再構築を防ぐ。最近の調査では、出力の切り捨てとDP保証が実用的な防御策として強調されている。

### #10.4.1 レベル: 1 役割: D

機微な属性が直接出力されないことを確認してください。必要に応じて、バケット化や一方向変換を使用してください。

### #10.4.2 レベル: 1 役割: D/V

同じ主体からの繰り返される適応クエリがクエリレート制限によって制御されていることを確認してください。

### #10.4.3 レベル: 2 役割: D

モデルがプライバシー保護ノイズで訓練されていることを検証してください。

## 10.5 モデル抽出防御

不正なクローン作成を検出し防止します。ウォーターマーキングとクエリパターン分析が推奨されます。

### #10.5.1 レベル: 1 役割: D

推論ゲートウェイが、モデルの記憶閾値に調整されたグローバルおよびAPIキーごとのレート制限を適用していることを検証してください。

### #10.5.2 レベル: 2 役割: D/V

問い合わせエントロピーと入力複数性の統計が自動抽出検出器に供給されていることを確認してください。

### #10.5.3 レベル: 2 役割: V

疑わしいクローンに対して1,000回以下のクエリで  $p < 0.01$  の確率で証明できる脆弱または確率的な透かしがあることを検証してください。

**#10.5.4 レベル: 3 役割: D**

ウォーターマークキーとトリガーセットがハードウェアセキュリティモジュールに保存され、毎年ローテーションされていることを検証してください。

**#10.5.5 レベル: 3 役割: V**

抽出アラートイベントに問題のあるクエリが含まれていることを確認し、それらがインシデント対応プレイブックと統合されていることを検証してください。

## 10.6 推論時の毒されたデータ検出

バックドアや毒された入力を特定し、中和します。

**#10.6.1 レベル: 1 役割: D**

モデル推論の前に、入力が異常検出器(例:STRIP、一貫性スコアリング)を通過することを確認してください。

**#10.6.2 レベル: 1 役割: V**

検出器の閾値が、5%未満の偽陽性を達成するために、クリーンおよび毒された検証セットで調整されていることを確認してください。

**#10.6.3 レベル: 2 役割: D**

汚染されたとフラグが付けられた入力がソフトブロッキングおよび人間によるレビューのワークフローを引き起こすことを確認します。

**#10.6.4 レベル: 2 役割: V**

検出器が適応型のトリガーレスバックドア攻撃でストレステストされていることを確認してください。

**#10.6.5 レベル: 3 役割: D**

検出効果指標がログに記録され、新しい脅威インテリジェンスと共に定期的に再評価されていることを確認してください。

## 10.7 動的セキュリティポリシー適応

脅威インテリジェンスと行動分析に基づくリアルタイムのセキュリティポリシー更新。

**#10.7.1 レベル: 1 役割: D/V**

エージェントの再起動なしにセキュリティポリシーを動的に更新でき、かつポリシーバージョンの整合性が維持されることを検証してください。

**#10.7.2 レベル: 2 役割: D/V**

ポリシーの更新が認可されたセキュリティ担当者によって暗号学的に署名され、適用前に検証されていることを確認してください。

**#10.7.3 レベル: 2 役割: D/V**

動的ポリシーの変更が、理由の説明、承認連鎖、およびロールバック手順を含む完全な監査記録と共に記録されていることを確認してください。

**#10.7.4 レベル: 3 役割: D/V**

適応型セキュリティメカニズムがリスクの文脈および行動パターンに基づいて脅威検出の感度を調整することを検証してください。

**#10.7.5 レベル: 3 役割: D/V**

ポリシー適応の決定が説明可能であり、セキュリティチームのレビューのために証拠のトレイルが含まれてい

ることを確認してください。

## 10.8 リフレクションベースのセキュリティ分析

エージェントの自己反省とメタ認知分析によるセキュリティ検証。

### #10.8.1 レベル: 1 役割: D/V

エージェントの反省メカニズムには、意思決定と行動に対するセキュリティに重点を置いた自己評価が含まれていることを確認してください。

### #10.8.2 レベル: 2 役割: D/V

自己評価メカニズムの操作を防ぐために、反射出力が検証されていることを確認してください。

### #10.8.3 レベル: 2 役割: D/V

メタ認知的セキュリティ分析が、エージェントの推論プロセスにおける潜在的なバイアス、操作、または妥協を特定することを確認してください。

### #10.8.4 レベル: 3 役割: D/V

リフレクションベースのセキュリティ警告が強化された監視および潜在的な人間の介入ワークフローをトリガーすることを確認してください。

### #10.8.5 レベル: 3 役割: D/V

セキュリティの振り返りからの継続的な学習が、正当な機能を損なうことなく脅威検出を改善することを検証する。

## 10.9 進化と自己改善のセキュリティ

自己改変および進化が可能なエージェントシステムのためのセキュリティ制御。

### #10.9.1 レベル: 1 役割: D/V

自己修正機能が指定された安全な領域内に制限されていることを、形式的検証の境界を用いて確認してください。

### #10.9.2 レベル: 2 役割: D/V

進化提案が実施される前にセキュリティ影響評価を受けることを確認してください。

### #10.9.3 レベル: 2 役割: D/V

自己改善メカニズムにおいて、整合性検証を伴うロールバック機能が含まれていることを確認してください。

### #10.9.4 レベル: 3 役割: D/V

メタラーニングのセキュリティが改善アルゴリズムの敵対的操作を防止することを検証する。

### #10.9.5 レベル: 3 役割: D/V

再帰的な自己改善が数学的収束の証明を伴う形式的安全制約によって制限されていることを検証する。

## 参考文献

- MITRE ATLAS adversary tactics for ML

- NIST AI Risk Management Framework 1.0, 2023
- OWASP Top 10 for LLM Applications, 2025
- Adversarial Training: A Survey — Zhao et al., 2024
- RobustBench adversarial-robustness benchmark
- Membership-Inference & Model-Inversion Risk Survey, 2025
- PURIFIER: Confidence-Score Defense against MI Attacks — AAAI 2023
- Model-Inversion Attacks & Defenses Survey — AI Review, 2025
- Comprehensive Defense Framework Against Model Extraction — IEEE TDSC 2024
- Fragile Model Watermarking Survey — 2025
- Data Poisoning in Deep Learning: A Survey — Zhao et al., 2025
- BDetCLIP: Multimodal Prompting Backdoor Detection — Niu et al., 2024



# 11 プライバシー保護と個人データ管理

## 制御目標

収集、トレーニング、推論、インシデント対応といったAIのライフサイクル全体にわたり厳格なプライバシー確保を維持し、個人データが明確な同意のもとで、必要最小限の範囲で、証明可能な消去および正式なプライバシー保証に基づいてのみ処理されるようにする。

### 11.1 匿名化とデータ最小化

#### #11.1.1 レベル: 1 役割: D/V

直接識別子および準識別子が削除またはハッシュ化されていることを確認してください。

#### #11.1.2 レベル: 2 役割: D/V

自動化された監査が k-匿名性 / l-多様性を測定し、それらの閾値がポリシーを下回った場合に警告を出すことを確認してください。

#### #11.1.3 レベル: 2 役割: V

モデルの特徴重要度レポートが、 $\epsilon = 0.01$  の相互情報量を超える識別子の漏洩がないことを証明していることを確認してください。

#### #11.1.4 レベル: 3 役割: V

形式的な証明や合成データの認証により、リンク攻撃が行われた場合でも再識別リスクが 0.05 以下であることを確認してください。

### 11.2 忘れられる権利と削除の強制

#### #11.2.1 レベル: 1 役割: D/V

データ主体の削除リクエストが、サービスレベルアグリーメント(SLA)内の30日未満で、生データセット、チェックポイント、埋め込み、ログ、およびバックアップにまで伝播することを検証してください。

#### #11.2.2 レベル: 2 役割: D

「マシンアンラーニング」ルーチンが、認定されたアンラーニングアルゴリズムを使用して物理的に再訓練または近似的な除去を行っていることを検証してください。

#### #11.2.3 レベル: 2 役割: V

シャドウモデル評価により、忘却後の出力に対する忘却対象レコードの影響が 1% 未満であることを検証してください。

#### #11.2.4 レベル: 3 役割: V

削除イベントが不変的に記録され、規制当局が監査可能であることを確認してください。

### 11.3 差分プライバシーの保護措置

#### #11.3.1 レベル: 2 役割: D/V

累積  $\epsilon$  がポリシー閾値を超えた場合にプライバシー損失会計ダッシュボードが警告を発することを検証する。

#### #11.3.2 レベル: 2 役割: V

ブラックボックスプライバシー監査が宣言された値の10%以内で  $\epsilon$  を推定していることを確認してください。

#### #11.3.3 レベル: 3 役割: V

形式的な証明が、トレーニング後のすべてのファインチューンおよび埋め込みを網羅していることを確認してください。

## 11.4 目的制限およびスコープクリープ保護

#### #11.4.1 レベル: 1 役割: D

すべてのデータセットおよびモデルチェックポイントに、元の同意内容に沿った機械可読の目的タグが付与されていることを確認してください。

#### #11.4.2 レベル: 1 役割: D/V

ランタイムモニターが、宣言された目的と矛盾するクエリを検出し、ソフト拒否をトリガーすることを確認する。

#### #11.4.3 レベル: 3 役割: D

ポリシー・アズ・コードのゲートが、DPIAレビューなしにモデルの新しいドメインへの再展開をブロックすることを検証してください。

#### #11.4.4 レベル: 3 役割: V

正式なトレースアビリティ証明が、すべての個人データのライフサイクルが同意された範囲内に留まっていることを示していることを検証します。

## 11.5 同意管理と合法的根拠のトラッキング

#### #11.5.1 レベル: 1 役割: D/V

同意管理プラットフォーム(CMP)が、データ主体ごとにオプトインの状態、目的、および保持期間を記録していることを確認してください。

#### #11.5.2 レベル: 2 役割: D

APIが同意トークンを公開していることを確認し、モデルは推論前にトークンスコープを検証する必要があります。

#### #11.5.3 レベル: 2 役割: D/V

拒否または撤回された同意が24時間以内に処理パイプラインを停止することを確認してください。

## 11.6 プライバシー制御を伴う連合学習

#### #11.6.1 レベル: 1 役割: D

クライアントの更新が集約前にローカル差分プライバシーのノイズ付加を適用していることを検証してください。

#### #11.6.2 レベル: 2 役割: D/V

トレーニングメトリクスが差分プライバシーを保持し、单一クライアントの損失を決して漏らさないことを検証してください。

#### #11.6.3 レベル: 2 役割: V

中毒耐性のある集約(例:Krum／Trimmed-Mean)が有効になっていることを確認してください。

#### #11.6.4 レベル: 3 役割: V

形式的な証明により、全体の  $\epsilon$  予算が5未満の効用損失であることを確認してください。

## 参考文献

- GDPR & AI Compliance Best Practices
- EU Parliament Study on GDPR & AI, 2020
- ISO 31700-1:2023 — Privacy by Design for Consumer Products
- NIST Privacy Framework 1.1 (2025 Draft)
- Machine Unlearning: Right-to-Be-Forgotten Techniques
- A Survey of Machine Unlearning, 2024
- Auditing DP-SGD — ArXiv 2024
- DP-SGD Explained — PyTorch Blog
- Purpose-Limitation for AI — IJLIT 2025
- Data-Protection Considerations for AI — URM Consulting
- Top Consent-Management Platforms, 2025
- Secure Aggregation in DP Federated Learning — ArXiv 2024

# C12 監視、ログ記録、および異常検知

## 制御目標

このセクションでは、モデルやその他のAIコンポーネントが何を見て、何をし、何を返すかについてリアルタイムおよびフォレンジックの可視性を提供するための要件を示しており、これにより脅威を検出、識別、学習することができます。

### C12.1 要求および応答のログ記録

#### #12.1.1 レベル: 1 役割: D/V

すべてのユーザー入力とモデルの応答が、適切なメタデータ(例: タイムスタンプ、ユーザーID、セッションID、モデルバージョン)と共に記録されていることを確認してください。

#### #12.1.2 レベル: 1 役割: D/V

ログが適切な保持ポリシーおよびバックアップ手順を備えた、安全でアクセス制御されたリポジトリに保存されていることを確認してください。

#### #12.1.3 レベル: 1 役割: D/V

ログストレージシステムが、ログに含まれる機密情報を保護するために保存時および転送時の暗号化を実装していることを確認してください。

#### #12.1.4 レベル: 1 役割: D/V

プロンプトおよび出力内の機密データがログ記録前に自動的に編集またはマスクされることを確認し、PII、認証情報、および機密情報に対する設定可能な編集ルールを適用してください。

#### #12.1.5 レベル: 2 役割: D/V

ポリシー決定および安全性フィルタリングの操作が、コンテンツモニタリングシステムの監査およびデバッグを可能にする十分な詳細で記録されていることを検証してください。

#### #12.1.6 レベル: 2 役割: D/V

ログの完全性が暗号署名や書き込み専用ストレージなどによって保護されていることを確認してください。

### C12.2 悪用検出および警告

#### #12.2.1 レベル: 1 役割: D/V

システムが既知のジャイルブレイクパターン、プロンプトインジェクションの試み、および署名ベースの検出を用いた敵対的入力を検出し、警告を発することを検証してください。

#### #12.2.2 レベル: 1 役割: D/V

システムが標準のログ形式およびプロトコルを使用して、既存のセキュリティ情報およびイベント管理(SIEM)プラットフォームと統合されていることを検証してください。

#### #12.2.3 レベル: 2 役割: D/V

強化されたセキュリティイベントに、モデル識別子、信頼度スコア、安全フィルタの判断など、AI特有のコンテキストが含まれていることを確認してください。

#### #12.2.4 レベル: 2 役割: D/V

行動異常検出が、異常な会話パターン、過剰な再試行試行、または体系的な探査行動を特定することを検証してください。

#### #12.2.5 レベル: 2 役割: D/V

潜在的なポリシー違反や攻撃試行が検出されたときに、リアルタイムのアラート機構がセキュリティチームに

通知することを検証してください。

#### #12.2.6 レベル: 2 役割: D/V

カスタムルールが、協調されたジャイルブレイク試行、プロンプトインジェクションキャンペーン、およびモデル抽出攻撃を含むAI特有の脅威パターンを検出するために含まれていることを確認してください。

#### #12.2.7 レベル: 3 役割: D/V

自動化されたインシデント対応ワークフローが、侵害されたモデルを隔離し、悪意のあるユーザーをロックし、重要なセキュリティイベントをエスカレーションできることを検証してください。

## C12.3 モデルドリフト検出

#### #12.3.1 レベル: 1 役割: D/V

システムがモデルのバージョンおよび時間経過にわたって精度、信頼スコア、レイテンシ、およびエラー率などの基本的なパフォーマンス指標を追跡していることを確認してください。

#### #12.3.2 レベル: 2 役割: D/V

性能指標が事前に定義された劣化閾値を超えるか、ベースラインから大きく逸脱した場合に自動アラートがトリガーされることを確認してください。

#### #12.3.3 レベル: 2 役割: D/V

モデル出力に事実誤認、一貫性のない情報、または虚構の情報が含まれている場合に、幻覚検出モニターがそれらの事例を識別してフラグを立てることを確認してください。

## C12.4 パフォーマンスおよび動作のテレメトリ

#### #12.4.1 レベル: 1 役割: D/V

リクエストの遅延、トーケン消費量、メモリ使用量、スループットを含む運用指標が継続的に収集および監視されていることを確認してください。

#### #12.4.2 レベル: 1 役割: D/V

成功率と失敗率が、エラータイプとその根本原因の分類とともに追跡されていることを確認してください。

#### #12.4.3 レベル: 2 役割: D/V

リソース利用状況の監視にGPU/CPU使用率、メモリ消費、ストレージ要件が含まれ、しきい値を超えた場合のアラートがあることを確認してください。

## C12.5 AI インシデント対応計画および実行

#### #12.5.1 レベル: 1 役割: D/V

インシデント対応計画が、モデルの損害、データポイズニング、敵対的攻撃を含むAI関連のセキュリティ事象に具体的に対応していることを確認してください。

#### #12.5.2 レベル: 2 役割: D/V

インシデント対応チームが、モデルの挙動や攻撃ベクターを調査するためのAI特有のフォレンジックツールと専門知識にアクセスできることを確認してください。

#### #12.5.3 レベル: 3 役割: D/V

インシデント後の分析に、モデル再訓練の検討、安全フィルターの更新、および教訓のセキュリティ管理への

統合が含まれていることを確認してください。

## C12.6 AIパフォーマンス劣化検出

AIモデルの性能と品質の劣化を時間経過で監視し検出する。

### #12.6.1 レベル: 1 役割: D/V

モデルの精度、適合率、再現率、およびF1スコアが継続的に監視され、ベースラインの閾値と比較されていることを確認してください。

### #12.6.2 レベル: 1 役割: D/V

データドリフト検出がモデルの性能に影響を与える可能性のある入力分布の変化を監視することを確認してください。

### #12.6.3 レベル: 2 役割: D/V

コンセプトドリフト検出が、入力と期待される出力の関係の変化を識別することを確認してください。

### #12.6.4 レベル: 2 役割: D/V

パフォーマンスの低下が自動アラートをトリガーし、モデルの再トレーニングまたは置換ワークフローを開始することを確認してください。

### #12.6.5 レベル: 3 役割: V

劣化の根本原因分析が、性能低下をデータの変更、インフラストラクチャの問題、または外部要因と関連付けていることを確認してください。

## C12.7 DAG 可視化とワークフローセキュリティ

ワークフロービジュアライゼーションシステムを情報漏洩および改ざん攻撃から保護する。

### #12.7.1 レベル: 1 役割: D/V

DAGの可視化データが保存または送信される前に、機密情報が除去されていることを確認してください。

### #12.7.2 レベル: 1 役割: D/V

ワークフローの可視化アクセス制御が、許可されたユーザーのみがエージェントの意思決定経路および推論のトレースを閲覧できることを確認してください。

### #12.7.3 レベル: 2 役割: D/V

DAGデータの整合性が暗号署名および改ざん検知ストレージ機構によって保護されていることを検証します。

### #12.7.4 レベル: 2 役割: D/V

ワークフロービジュアライゼーションシステムが、細工されたノードまたはエッジデータを介したインジェクション攻撃を防ぐために入力検証を実装していることを確認してください。

### #12.7.5 レベル: 3 役割: D/V

リアルタイムのDAG更新がレート制限され、可視化システムへのサービス拒否攻撃を防ぐために検証されていることを確認してください。

## C12.8 プロアクティブなセキュリティ行動モニタリング

プロアクティブなエージェント行動分析によるセキュリティ脅威の検出と防止。

#### #12.8.1 レベル: 1 役割: D/V

リスク評価の統合とともに、プロアクティブなエージェントの動作が実行前にセキュリティ検証されていることを確認してください。

#### #12.8.2 レベル: 2 役割: D/V

自律的イニシアティブのトリガーには、セキュリティコンテキストの評価と脅威の状況分析が含まれていることを確認してください。

#### #12.8.3 レベル: 2 役割: D/V

潜在的なセキュリティへの影響や意図しない結果について、プロアクティブな行動パターンが分析されていることを確認してください。

#### #12.8.4 レベル: 3 役割: D/V

セキュリティに重要な先制的な行動は、監査履歴を伴う明示的な承認チェーンを必要とするかを確認してください。

#### #12.8.5 レベル: 3 役割: D/V

行動異常検知が、妥協を示す可能性のあるプロアクティブエージェントのパターンの逸脱を識別することを検証する。

## 参考文献

- NIST AI Risk Management Framework 1.0 - Manage 4.1 and 4.3
- ISO/IEC 42001:2023 – AI Management Systems Requirements - Annex B 6.2.6

# C13 人間の監督、説明責任、およびガバナンス

## 制御目標

この章では、AIシステムにおける人間の監督と明確な責任の連鎖を維持するための要件を提供し、AIのライフサイクル全体にわたって説明可能性、透明性、および倫理的管理を確保します。

### C13.1 キルスイッチおよびオーバーライドメカニズム

AIシステムの安全でない動作が観察された場合に備え、シャットダウンまたはロールバックの経路を提供してください。

#### #13.1.1 レベル: 1 役割: D/V

AIモデルの推論および出力を即座に停止するための手動キルスイッチ機構が存在することを確認してください。

#### #13.1.2 レベル: 1 役割: D

オーバーライドコントロールが認可された担当者のみにアクセス可能であることを確認してください。

#### #13.1.3 レベル: 3 役割: D/V

ロールバック手順が以前のモデルバージョンまたはセーフモード操作に戻せることを検証してください。

#### #13.1.4 レベル: 3 役割: V

オーバーライドメカニズムが定期的にテストされていることを確認してください。

### C13.2 ヒューマン・イン・ザ・ループ意思決定チェックポイント

リスクの事前定義された閾値を超えた場合に人間の承認を必要とする。

#### #13.2.1 レベル: 1 役割: D/V

高リスクなAIの判断は、実行前に明確な人間の承認が必要であることを確認してください。

#### #13.2.2 レベル: 1 役割: D

リスク閾値が明確に定義されており、自動的に人間によるレビューのワークフローをトリガーすることを確認してください。

#### #13.2.3 レベル: 2 役割: D

必要な時間内に人間の承認が得られない場合に備えて、時間に敏感な決定にはフォールバック手順があることを確認してください。

#### #13.2.4 レベル: 3 役割: D/V

該当する場合、エスカレーション手続きが異なる意思決定タイプまたはリスクカテゴリに対して明確な権限レベルを定義していることを確認してください。

## C13.3 責任の連鎖と監査可能性

オペレーターの操作とモデルの判断を記録する。

### #13.3.1 レベル: 1 役割: D/V

すべてのAIシステムの意思決定および人間の介入が、タイムスタンプ、ユーザーID、および意思決定の根拠とともに記録されていることを確認してください。

### #13.3.2 レベル: 2 役割: D

監査ログが改ざんできないことを確認し、完全性検証の仕組みを含めること。

## C13.4 説明可能なAI技術

表面特徴の重要性、反実仮想、および局所的説明。

### #13.4.1 レベル: 1 役割: D/V

AIシステムがその決定に対して人間が読める形式で基本的な説明を提供していることを確認してください。

### #13.4.2 レベル: 2 役割: V

説明の質が人間の評価研究および指標によって検証されていることを確認してください。

### #13.4.3 レベル: 3 役割: D/V

重要な意思決定に対して、特徴重要度スコアや帰属方法(SHAP、LIMEなど)が利用可能であることを確認してください。

### #13.4.4 レベル: 3 役割: V

反事実説明が、該当するユースケースおよびドメインにおいて、入力をどのように変更すれば結果が変わるかを示していることを確認してください。

## C13.5 モデルカードと使用開示

モデルカードを、意図された使用法、性能指標、および倫理的考慮事項について管理する。

### #13.5.1 レベル: 1 役割: D

モデルカードが意図された使用例、制限、および既知の失敗モードを文書化していることを確認してください。

### #13.5.2 レベル: 1 役割: D/V

異なる適用可能なユースケースにおけるパフォーマンス指標が開示されていることを確認してください。

### #13.5.3 レベル: 2 役割: D

倫理的配慮、バイアス評価、公平性評価、トレーニングデータの特性、および既知のトレーニングデータの制限が文書化され、定期的に更新されていることを確認してください。

### #13.5.4 レベル: 2 役割: D/V

モデルカードがバージョン管理され、変更追跡とともにモデルのライフサイクル全体で維持されていることを確認してください。

## C13.6 不確実性の定量化

応答における信頼度スコアまたはエントロピー測定値を伝播させる。

### #13.6.1 レベル: 1 役割: D

AIシステムが出力とともに信頼度スコアまたは不確実性の測定値を提供していることを確認してください。

### #13.6.2 レベル: 2 役割: D/V

不確実性の閾値が追加の人間によるレビューや代替の意思決定経路を引き起こすことを確認してください。

### #13.6.3 レベル: 2 役割: V

不確実性の定量化手法が実測データに対して校正および検証されていることを確認してください。

### #13.6.4 レベル: 3 役割: D/V

不確実性の伝播が複数段階のAIワークフローを通じて維持されていることを検証する。

## C13.7 ユーザ向け透明性レポート

インシデント、ドリフト、およびデータ使用状況について定期的に開示を行う。

### #13.7.1 レベル: 1 役割: D/V

データ使用ポリシーおよびユーザー同意管理の実践が、利害関係者に明確に伝えられていることを確認してください。

### #13.7.2 レベル: 2 役割: D/V

AI影響評価が実施され、その結果が報告に含まれていることを確認してください。

### #13.7.3 レベル: 2 役割: D/V

定期的に公開される透明性レポートが、AIのインシデントおよび運用指標を適切な詳細で開示していることを確認してください。

## 参考文献

- EU Artificial Intelligence Act — Regulation (EU) 2024/1689 (Official Journal, 12 July 2024)
- ISO/IEC 23894:2023 — Artificial Intelligence — Guidance on Risk Management
- ISO/IEC 42001:2023 — AI Management Systems Requirements
- NIST AI Risk Management Framework 1.0
- NIST SP 800-53 Revision 5 — Security and Privacy Controls
- A Unified Approach to Interpreting Model Predictions (SHAP, ICML 2017)
- Model Cards for Model Reporting (Mitchell et al., 2018)
- Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (Gal & Ghahramani, 2016)
- ISO/IEC 24029-2:2023 — Robustness of Neural Networks — Methodology for Formal Methods
- IEEE 7001-2021 — Transparency of Autonomous Systems
- Human Oversight under Article 14 of the EU AI Act (Fink, 2025)

## 付録A:用語集

この包括的な用語集は、AISVS全体で使用される主要なAI、ML、およびセキュリティ用語の定義を提供し、明確さと共通理解を確保します。

- 敵対的サンプル: 人間には気づかれない微妙な挿動を加えることで、AIモデルに誤りを引き起こすよう意図的に作成された入力データ。
- 敵対的ロバスト性 – AIにおける敵対的ロバスト性とは、故意に作成された悪意のある入力によってエラーを引き起こされることなく、モデルがその性能を維持し続ける能力を指します。
- エージェント – AIエージェントは、ユーザーに代わって目標を追求し、タスクを完了するためにAIを使用するソフトウェアシステムです。これらは推論、計画、記憶を示し、意思決定、学習、適応を行う一定の自律性を持っています。
- エージェンシックAI: ある程度の自律性を持って目標を達成するために動作できるAIシステムであり、多くの場合、人間の直接的な介入なしに意思決定や行動を行います。
- 属性ベースアクセス制御(ABAC): ユーザー、リソース、アクション、および環境の属性に基づいて認可の判断を行い、クエリ時に評価されるアクセス制御のパラダイム。
- バックドア攻撃: 特定のトリガーに対してモデルが特定の反応を示すように訓練され、それ以外の場合は正常に動作するデータポイズニング攻撃の一種。
- バイアス: 特定のグループや特定の状況において不公平または差別的な結果を招く可能性のある、AIモデルの出力における体系的な誤り。
- バイアス悪用: AIモデルの既知のバイアスを利用して出力や結果を操作する攻撃手法。
- Cedar: AIシステムのABAC実装に使用される、細粒度の権限を定義するAmazonのポリシー言語およびエンジン。
- Chain of Thought: 最終的な回答を生成する前に中間の推論ステップを生成することで、言語モデルの推論能力を向上させる技術。
- サーキットブレーカー: 特定のリスク閾値を超えた場合に自動的にAIシステムの動作を停止するメカニズム。
- 機密推論サービス: 信頼できる実行環境(TEE)または同等の機密コンピューティング機構内でAIモデルを実行し、モデルの重みおよび推論データが暗号化され、封印され、無許可のアク

セスや改ざんから保護される推論サービス。

- ・機密ワークロード:コード、データ、およびモデルをホストや共同テナントのアクセスから保護するために、ハードウェアによる分離、メモリ暗号化、およびリモート認証を備えた信頼できる実行環境(TEE)内で実行されるAIワークロード(例:トレーニング、推論、前処理)。
- ・データリーク:AIモデルの出力や挙動を通じた機密情報の意図しない露出。
- ・データポイズニング:モデルの整合性を損なうためにトレーニングデータを意図的に破損されることであり、バックドアを仕込んだり性能を低下させたりすることが多い。
- ・差分プライバシー – 差分プライバシーは、個々のデータ対象者のプライバシーを保護しながらデータセットに関する統計情報を公開するための数学的に厳密なフレームワークです。これにより、データ保有者は個別の個人に関する情報の漏洩を制限しつつ、グループの集約的なパターンを共有することができます。
- ・埋め込み:高次元空間において意味的な意味を捉えるデータ(テキスト、画像など)の密なベクトル表現。
- ・説明可能性 – AIにおける説明可能性とは、AIシステムがその意思決定や予測に対して人間が理解できる理由を提供し、その内部の仕組みに関する洞察を示す能力のことです。
- ・説明可能なAI(XAI):さまざまな技術やフレームワークを通じて、意思決定や行動について人間が理解できる説明を提供するように設計されたAIシステム。
- ・フェデレーテッドラーニング:データ自体を交換することなく、複数の分散デバイス上のローカルデータサンプルを保持し、それらにまたがってモデルを訓練する機械学習の手法。
- ・定式化:ハイパーパラメータ、トレーニング構成、前処理ステップ、ビルドスクリプトなど、アーティファクトやデータセットを生成するために使用されるレシピまたは方法。
- ・ガードレール:AIシステムが有害、偏った、またはその他望ましくない出力を生成するのを防ぐために実装される制約。
- ・幻覚 – AIの幻覚とは、AIモデルが訓練データや事実に基づかない誤ったまたは誤解を招く情報を生成する現象を指します。
- ・ヒューマン・イン・ザ・ループ(HITL):重要な意思決定のポイントで人間の監視、検証、または介入を必要とするよう設計されたシステム。

- Infrastructure as Code (IaC): 手動プロセスの代わりにコードを通じてインフラストラクチャを管理およびプロビジョニングし、セキュリティスキヤンと一貫したデプロイメントを可能にします。
- Jailbreak: 特に大規模言語モデルにおいて、AIシステムの安全ガードレールを回避し、禁止されたコンテンツを生成するために使用される手法。
- 最小権限: ユーザーやプロセスに対して必要最小限のアクセス権のみを付与するセキュリティ原則。
- LIME(局所的に解釈可能なモデル非依存型説明): 任意の機械学習分類器の予測を、解釈可能なモデルで局所的に近似することにより説明する手法。
- MCP(モデルコンテキストプロトコル): AIモデルやエージェントが、構造化された型付きのリクエストおよびレスポンスを定義されたトランスポート経由で交換することにより、外部ツール、データソース、およびリソースにアクセスできるプロトコル。
- メンバーシップ推論攻撃: 特定のデータポイントが機械学習モデルの訓練に使用されたかどうかを判定することを目的とした攻撃。
- MITRE ATLAS: 人工知能システムに対する敵対的脅威の全体像; AIシステムに対する敵対的な戦術と技術のナレッジベース。
- モデルカード – モデルカードとは、AIモデルの性能、制限、意図された使用法、および倫理的配慮に関する標準化された情報を提供し、透明性と責任あるAI開発を促進する文書です。
- モデル抽出: 攻撃者が対象モデルに繰り返しクエリを送り、無許可で機能的に類似したコピーを作成する攻撃。
- モデル反転攻撃: モデルの出力を解析することで訓練データを再構築しようとする攻撃。
- モデルライフサイクル管理 – AIモデルライフサイクル管理とは、AIモデルの設計、開発、展開、監視、保守、および最終的な廃止を含むすべての段階を監督し、モデルが効果的で目的に適合し続けるようにするプロセスです。
- モデルポイズニング: トレーニングプロセス中にモデルに直接脆弱性やバックドアを導入すること。
- モデル窃盗/盗用: 繰り返しのクエリを通じて、独自モデルのコピーまたは近似を抽出すること。

- マルチエージェントシステム:複数の相互作用するAIエージェントで構成され、それぞれが異なる能力や目標を持つ可能性のあるシステム。
- OPA(Open Policy Agent):スタック全体で統一されたポリシー適用を可能にするオープンソースのポリシーエンジン。
- プライバシー保護機械学習(PPML):トレーニングデータのプライバシーを保護しながら、機械学習モデルを訓練および展開するための技術と方法。
- プロンプトインジェクション:モデルの意図した動作を上書きするために、悪意のある指示が入力に埋め込まれる攻撃。
- RAG(Retrieval-Augmented Generation):応答を生成する前に外部の知識ソースから関連情報を取得することで、大規模言語モデルを強化する技術。
- レッドチームинг:敵対的攻撃をシミュレートしてAIシステムを積極的にテストし、脆弱性を特定する手法。
- SBOM(ソフトウェア部品表):ソフトウェアやAIモデルの構築に使用されるさまざまなコンポーネントの詳細およびサプライチェーンの関係を含む正式な記録。
- SHAP(SHapley Additive exPlanations):各特徴量の予測への寄与度を計算することにより、任意の機械学習モデルの出力を説明するゲーム理論に基づく手法。
- 強力な認証:少なくとも二要素(知識、所持、固有)を要求し、FIDO2/WebAuthn、証明書ベースのサービス認証、または短期間有効なトークンなどのフィッシング耐性メカニズムを用いることで、資格情報の窃盗やリプレイ攻撃に対抗する認証方式。
- サプライチェーン攻撃:サードパーティのライブラリ、データセット、または事前学習済みモデルなど、供給連鎖における安全性の低い要素を攻撃することによってシステムを侵害すること。
- 転移学習:あるタスクのために開発されたモデルを、別のタスクのモデルの出発点として再利用する技術。
- ベクターデータベース:高次元ベクトル(埋め込み)を保存し、効率的な類似検索を実行するために設計された専門的なデータベース。
- 脆弱性スキャン:AIフレームワークや依存関係を含むソフトウェアコンポーネントに存在する既知のセキュリティ脆弱性を特定する自動化ツール。

- ・ウォーターマーキング:AI生成コンテンツに目に見えないマーカーを埋め込み、その起源を追跡したりAI生成を検出したりする技術。
- ・ゼロデイ脆弱性:開発者がパッチを作成および配布する前に、攻撃者が悪用できる以前は知られていなかった脆弱性。



## 付録B: 参考文献

### 未完了事項



# 付録C:AIセキュリティガバナンスと文書化(再編成)

## 目的

この付録は、システムライフサイクル全体でAIセキュリティを管理するための組織構造、ポリシー、文書、およびプロセスを確立するための基本要件を提供します。

## AC.1 AIリスク管理フレームワークの採用

### #AC.1.1 レベル: 1 役割: D/V

AI特有のリスク評価方法論が文書化され、実施されていることを確認してください。

### #AC.1.2 レベル: 2 役割: D

AIライフサイクルの重要なポイントおよび重大な変更の前にリスク評価が実施されていることを確認してください。

### #AC.1.3 レベル: 3 役割: D/V

リスク管理フレームワークが確立された標準(例:NIST AI RMF)と整合していることを確認する。

## AC.2 AIセキュリティポリシーと手順

### #AC.2.1 レベル: 1 役割: D/V

文書化されたAIセキュリティポリシーが存在することを確認してください。

### #AC.2.2 レベル: 2 役割: D

ポリシーが少なくとも年に一度、および重大な脅威環境の変化の後に見直され、更新されていることを確認してください。

### #AC.2.3 レベル: 3 役割: D/V

ポリシーがすべてのAISVSカテゴリおよび適用される規制要件に対応していることを確認してください。

## AC.3 AIセキュリティの役割と責任

### #AC.3.1 レベル: 1 役割: D/V

AIのセキュリティロールと責任が文書化されていることを確認してください。

### #AC.3.2 レベル: 2 役割: D

責任者が適切なセキュリティ専門知識を有していることを確認してください。

### #AC.3.3 レベル: 3 役割: D/V

高リスクAIシステムに対して、AI倫理委員会またはガバナンス委員会が設立されていることを確認してください。

## AC.4 倫理的なAIガイドラインの施行

### #AC.4.1 レベル: 1 役割: D/V

AI開発および展開に関する倫理ガイドラインが存在することを確認してください。

### #AC.4.2 レベル: 2 役割: D

倫理違反を検出し報告するための仕組みが整備されていることを確認してください。

### #AC.4.3 レベル: 3 役割: D/V

展開されたAIシステムに対して定期的な倫理レビューが実施されていることを確認してください。

## AC.5 AI 規制遵守モニタリング

### #AC.5.1 レベル: 1 役割: D/V

適用されるAI規制を特定するためのプロセスが存在することを確認してください。

### #AC.5.2 レベル: 2 役割: D

すべての規制要件への準拠が評価されていることを確認してください。

### #AC.5.3 レベル: 3 役割: D/V

規制の変更がAIシステムの適時なレビューおよび更新を引き起こすことを確認してください。

## AC.6 トレーニングデータのガバナンス、ドキュメント化とプロセス

### AC.6.1 データソーシングとデューデリジェンス

#### #AC.6.1.1 レベル: 1 役割: D/V

品質、代表性、倫理的な調達、およびライセンス遵守が検証されたデータセットのみを許可し、毒性、埋め込まれたバイアス、知的財産権侵害のリスクを低減することを確認してください。

#### #AC.6.1.2 レベル: 2 役割: D/V

事前に第三者データ供給者(事前学習済みモデルの提供者や外部データセットの提供者を含む)が、データやモデルが統合される前に、セキュリティ、プライバシー、倫理的調達、およびデータ品質のデューデリジェンスを受けていることを確認する。

#### #AC.6.1.3 レベル: 1 役割: D

外部転送がTLS/認証および整合性チェックを使用していることを確認してください。

#### #AC.6.1.4 レベル: 2 役割: D/V

高リスクのデータソース(例:出所不明のオープンソースデータセット、検証されていないサプライヤー)が、機微なアプリケーションで使用される前に、サンドボックス分析、徹底した品質・バイアスチェック、標的型汚染検出などの強化された精査を受けていることを確認してください。

#### #AC.6.1.5 レベル: 3 役割: D/V

第三者から取得した事前学習済みモデルが、微調整や展開の前に、埋め込まれたバイアス、潜在的なバックドア、アーキテクチャの完全性、および元のトレーニングデータの出所について評価されていることを確認してください。

### AC.6.2 バイアスと公平性の管理

**#AC.6.2.1 レベル: 1 役割: D/V**

データセットが、法的に保護された属性(例:人種、性別、年齢)およびモデルの適用領域に関連するその他の倫理的に敏感な特性(例:社会経済的地位、所在地)における代表性の不均衡や潜在的なバイアスについてプロファイリングされていることを検証する。

**#AC.6.2.2 レベル: 2 役割: D/V**

特定されたバイアスが、再バランス、ターゲットを絞ったデータ拡張、アルゴリズム調整(例:前処理、処理中、後処理技術)、または再重み付けなどの文書化された戦略によって軽減されていることを検証し、軽減措置が公平性および全体的なモデル性能の両方に与える影響が評価されていることを確認してください。

**#AC.6.2.3 レベル: 2 役割: D/V**

トレーニング後の公平性メトリクスが評価され、文書化されていることを確認してください。

**#AC.6.2.4 レベル: 3 役割: D/V**

ライフサイクルバイアス管理ポリシーが所有者とレビューの周期を割り当てていることを確認してください。

## AC.6.3 ラベリングおよび注釈のガバナンス

**#AC.6.3.1 レベル: 2 役割: D/V**

ラベリング／アノテーションの品質がレビューのクロスチェックや合意形成を通じて確保されていることを確認してください。

**#AC.6.3.2 レベル: 2 役割: D/V**

重要なトレーニングデータセットについて、特徴、目的、構成、収集プロセス、前処理、ライセンス、および推奨される使用法・非推奨使用法の詳細を記載したデータカードが維持されていることを確認してください。

**#AC.6.3.3 レベル: 2 役割: D/V**

データカードが、データセットに関連するバイアスのリスク、人口統計の偏り、および倫理的考慮事項を文書化していることを検証してください。

**#AC.6.3.4 レベル: 2 役割: D/V**

データカードがデータセットと共にバージョン管理され、データセットが変更されるたびに更新されていることを確認してください。

**#AC.6.3.5 レベル: 2 役割: D/V**

データカードが技術関係者および非技術関係者(例:コンプライアンス、倫理、ドメインエキスパート)の双方によってレビューおよび承認されていることを確認してください。

**#AC.6.3.6 レベル: 2 役割: D/V**

ラベリング／アノテーションの品質が、明確なガイドライン、レビューによる相互チェック、合意形成メカニズム(例:アノテーター間一致度の監視)、および不一致を解決するための定義されたプロセスを通じて確保されていることを確認する。

**#AC.6.3.7 レベル: 3 役割: D/V**

安全性、セキュリティ、公平性にとって重要なラベル(例:有害なコンテンツの識別、重要な医療所見)が、必須の独立二重レビューまたは同等の厳格な検証を受けていることを確認してください。

**#AC.6.3.8 レベル: 2 役割: D/V**

ラベリングガイドと指示が包括的で、バージョン管理され、ピアレビューを受けていることを確認してください。

**#AC.6.3.9 レベル: 2 役割: D/V**

ラベルのデータスキームが明確に定義されており、バージョン管理されていることを確認してください。

**#AC.6.3.10 レベル: 2 役割: D/V**

アウトソーシングまたはクラウドソーシングされたラベリングワークフローに、データの機密性、完全性、ラベル品質を確保し、データ漏洩を防止するための技術的および手続き的な安全対策が含まれていることを確認してください。

**#AC.6.3.11 レベル: 2 役割: D/V**

データ注釈に関する全ての担当者が、身元調査を受け、データセキュリティおよびプライバシーに関する訓練を受けていることを確認してください。

**#AC.6.3.12 レベル: 2 役割: D/V**

すべてのアノテーション担当者が機密保持及び秘密保持契約に署名していることを確認してください。

#### #AC.6.3.13 レベル: 2 役割: D/V

注釈プラットフォームがアクセス制御を実施し、内部脅威を監視していることを確認する。

### AC.6.4 データセット品質ゲートおよび検疫

#### #AC.6.4.1 レベル: 2 役割: D/V

失敗したデータセットが監査証跡とともに隔離されていることを確認してください。

#### #AC.6.4.2 レベル: 2 役割: D/V

品質ゲートが例外が承認されない限り、基準に満たないデータセットをブロックすることを確認してください。

#### #AC.6.4.3 レベル: 2 役割: V

ドメイン専門家による手動スポットチェックが、統計的に有意なサンプル(例:1%以上または1,000サンプル以上のいずれか大きい方、またはリスク評価によって決定される)をカバーしており、自動化では検出できない微妙な品質問題を特定できることを確認してください。

### AC.6.5 脅威/毒性検出およびドリフト

#### #AC.6.5.1 レベル: 2 役割: D/V

フラグが付けられたサンプルがトレーニング前に手動レビューを引き起こすことを確認してください。

#### #AC.6.5.2 レベル: 2 役割: V

結果がモデルのセキュリティドシエにフィードされ、継続的な脅威インテリジェンスに情報を提供していることを確認する。

#### #AC.6.5.3 レベル: 3 役割: D/V

検出口ジックが新しい脅威インテリジェンスで更新されていることを確認してください。

#### #AC.6.5.4 レベル: 3 役割: D/V

オンライン学習パイプラインが分布の変動を監視していることを確認してください。

### AC.6.6 削除、同意、権利、保持およびコンプライアンス

#### #AC.6.6.1 レベル: 1 役割: D/V

トレーニングデータの削除ワークフローが一次データおよび派生データを完全に削除し、モデルへの影響を評価していることを確認し、影響を受けたモデルに対して必要に応じて再トレーニングや再調整などの対処が行われていることを評価してください。

#### #AC.6.6.2 レベル: 2 役割: D

トレーニングに使用されるデータに対するユーザーの同意(および撤回)の範囲と状況を追跡・尊重するための仕組みが整っていることを確認し、新しいトレーニングプロセスや重要なモデル更新にデータが組み込まれる前に同意が検証されていることを確認してください。

#### #AC.6.6.3 レベル: 2 役割: V

ワークフローが毎年テストされ、記録されていることを確認してください。

#### #AC.6.6.4 レベル: 1 役割: D/V

すべてのトレーニングデータセットに対して、明示的な保持期間が定義されていることを確認してください。

#### #AC.6.6.5 レベル: 2 役割: D/V

データセットがライフサイクルの終了時に自動的に期限切れとなり、削除されるか、削除のためにレビューされることを確認してください。

#### #AC.6.6.6 レベル: 2 役割: D/V

保持および削除の操作が記録され、監査可能であることを確認してください。

#### #AC.6.6.7 レベル: 2 役割: D/V

すべてのデータセットに対して、データ所在と国境を越えた転送の要件が特定され、適用されていることを確

認してください。

#### #AC.6.6.8 レベル: 2 役割: D/V

データ処理において、業界特有の規制(例:医療、金融)が特定され対処されていることを確認する。

#### #AC.6.6.9 レベル: 2 役割: D/V

関連するプライバシー法(例:GDPR、CCPA)への準拠が文書化され、定期的に見直されていることを確認してください。

#### #AC.6.6.10 レベル: 2 役割: D/V

データ主体のアクセス、訂正、制限、または異議申し立ての要求に対応するための仕組みが存在することを確認してください。

#### #AC.6.6.11 レベル: 2 役割: D/V

リクエストが法的に定められた期限内に記録され、追跡され、履行されていることを確認してください。

#### #AC.6.6.12 レベル: 2 役割: D/V

データ主体の権利に関するプロセスが効果的であるかどうかを定期的にテストおよびレビューすることを確認してください。

### AC.6.7 バージョニングおよび変更管理

#### #AC.6.7.1 レベル: 2 役割: D/V

データセットバージョンを更新または置換する前に、モデルのパフォーマンス、公平性、およびコンプライアンスを含むインパクト分析が実施されていることを検証してください。

#### #AC.6.7.2 レベル: 2 役割: D/V

影響分析の結果が文書化され、関連する利害関係者によってレビューされていることを確認する。

#### #AC.6.7.3 レベル: 2 役割: D/V

新しいバージョンが許容できないリスクや後退をもたらした場合に備え、ロールバックプランが存在することを確認してください。

### AC.6.8 合成データガバナンス

#### #AC.6.8.1 レベル: 2 役割: D/V

生成プロセス、パラメーター、および合成データの意図された使用が文書化されていることを確認してください。

#### #AC.6.8.2 レベル: 2 役割: D/V

合成データをトレーニングに使用する前に、バイアス、プライバシー漏洩、および表現上の問題についてリスク評価が行われていることを確認してください。

### AC.6.9 アクセスマニタリング

#### #AC.6.9.1 レベル: 2 役割: D/V

アクセスログが定期的にレビューされており、大量のエクスポートや新しい場所からのアクセスなどの異常なパターンが確認されていることを検証してください。

#### #AC.6.9.2 レベル: 2 役割: D/V

疑わしいアクセスイベントに対してアラートが生成され、速やかに調査されていることを確認してください。

### AC.6.10 敵対的トレーニングガバナンス

#### #AC.6.10.1 レベル: 2 役割: D/V

敵対的トレーニングが使用されている場合、敵対的データセットの生成、管理、およびバージョニングが文書

化され、管理されていることを確認してください。

#### #AC.6.10.2 レベル: 3 役割: D/V

敵対的ロバストネストレーニングがモデルの性能(クリーン入力および敵対的入力の両方に対して)および公平性指標に与える影響が評価され、文書化され、監視されていることを確認してください。

#### #AC.6.10.3 レベル: 3 役割: D/V

敵対的トレーニングと堅牢性のための戦略が、進化する敵対的攻撃技術に対抗するために定期的に見直され、更新されていることを確認してください。

## AC.7 モデルライフサイクルガバナンスとドキュメント管理

#### #AC.7.1 レベル: 2 役割: D/V

すべてのモデルアーティファクトがセマンティックバージョニング(MAJOR.MINOR.PATCH)を使用していることを確認し、各バージョンコンポーネントが増分される条件を文書化してください。

#### #AC.7.2 レベル: 2 役割: D/V

緊急展開には、文書化されたセキュリティリスク評価および事前に指定されたセキュリティ権限からの承認が、事前に合意された期間内に必要であることを確認してください。

#### #AC.7.3 レベル: 2 役割: V

ロールバックアーティファクト(前のモデルバージョン、構成、依存関係)が組織のポリシーに従って保持されていることを確認してください。

#### #AC.7.4 レベル: 2 役割: D/V

監査ログへのアクセスには適切な承認が必要であること、およびすべてのアクセス試行がユーザー識別情報とタイムスタンプとともに記録されることを確認してください。

#### #AC.7.5 レベル: 1 役割: D/V

退役モデルの成果物がデータ保持ポリシーに従って保持されていることを確認してください。

## AC.8 プロンプト、入力、および出力の安全性ガバナンス

### AC.8.1 プロンプトイインジェクション防御

#### #AC.8.1.1 レベル: 2 役割: D/V

敵対的評価テスト(例:レッドチームの「多回ショット」プロンプト)が、すべてのモデルまたはプロンプテンプレートのリリース前に実行され、成功率の閾値および回帰に対する自動ブロッカーが設定されていることを確認してください。

#### #AC.8.1.2 レベル: 3 役割: D/V

すべてのプロンプトフィルタールールの更新、分類モデルのバージョン、およびブロックリストの変更がバージョン管理され、監査可能であることを確認してください。

### AC.8.2 敵対的サンプル耐性

#### #AC.8.2.1 レベル: 3 役割: D/V

既知の攻撃スイートの成功率という堅牢性指標が、自動化によって時間経過とともに追跡され、回帰が検出された場合にアラートが発報されることを確認してください。

## AC.8.3 コンテンツおよびポリシースクリーニング

### #AC.8.3.1 レベル: 2 役割: D

スクリーニングモデルまたはルールセットが少なくとも四半期ごとに再訓練または更新され、新たに観察されたジャイルブレイクやポリシーバイパスのパターンが組み込まれていることを確認してください。

## AC.8.4 入力レート制限と悪用防止

### #AC.8.4.1 レベル: 3 役割: V

悪用防止ログが保持され、発生しつつある攻撃パターンのためにレビューされていることを確認してください。

## AC.8.5 入力の出所と帰属

### #AC.8.5.1 レベル: 1 役割: D/V

すべてのユーザー入力が取り込み時にメタデータ(ユーザーID、セッション、ソース、タイムスタンプ、IPアドレス)でタグ付けされていることを検証してください。

### #AC.8.5.2 レベル: 2 役割: D/V

すべての処理された入力に対して、プロヴィナンスマタデータが保持され監査可能であることを確認してください。

### #AC.8.5.3 レベル: 2 役割: D/V

異常または信頼できない入力ソースが検出され、強化された精査またはブロックの対象となることを確認してください。

## AC.9 マルチモーダル検証、MLOpsおよびインフラストラクチャガバナンス

### AC.9.1 マルチモーダルセキュリティ検証パイプライン

#### #AC.9.1.1 レベル: 3 役割: D/V

モダリティ固有のコンテンツ分類器が、文書化されたスケジュール(最低でも四半期ごと)に従って新しい脅威パターンや敵対的事例、性能ベンチマークとともに更新され、ベースラインの閾値を上回って維持されていることを検証してください。

### AC.9.2 CI/CDおよびビルドセキュリティ

#### #AC.9.2.1 レベル: 1 役割: D/V

インフラストラクチャ・アズ・コードがすべてのコミットでスキャンされていることを確認し、重大または高い重大度の検出結果がある場合はマージをブロックします。

#### #AC.9.2.2 レベル: 2 役割: D/V

CI/CDパイプラインがシークレットやインフラストラクチャへのアクセスに短命でスコープ限定のアイデンティティを使用していることを検証してください。

#### #AC.9.2.3 レベル: 2 役割: D/V

ビルド環境が本番ネットワークおよびデータから分離されていることを確認してください。

### AC.9.3 コンテナおよびイメージのセキュリティ

**#AC.9.3.1 レベル: 2 役割: D/V**

コンテナイメージがスキャンされ、ハードコーディングされたシークレット(例:APIキー、認証情報、証明書)をブロックしていることを確認してください。

**#AC.9.3.2 レベル: 1 役割: D/V**

コンテナイメージが組織のスケジュールに従ってスキャンされていること、および組織のリスク閾値に基づいてCRITICALな脆弱性が検出された場合に展開がブロックされることを確認してください。

## AC.9.4 監視、警報、およびSIEM

**#AC.9.4.1 レベル: 2 役割: V**

セキュリティアラートがCEFまたはSTIX/TAXIIフォーマットを使用し、自動化されたエンリッチメントを伴ってSIEMプラットフォーム(Splunk、Elastic、またはSentinel)と統合されていることを検証してください。

## AC.9.5 脆弱性管理

**#AC.9.5.1 レベル: 2 役割: D/V**

HIGHの深刻度を持つ脆弱性が、組織のリスク管理タイムラインに従って修正されていることを確認し、かつ、現在悪用されているCVEに対しては緊急対応手順が適用されていることを確認してください。

## AC.9.6 設定およびドリフト制御

**#AC.9.6.1 レベル: 2 役割: D/V**

組織の監視要件に従って、ツール(Chef InSpec、AWS Config)を使用して構成のドリフトが検出されることを確認し、許可されていない変更に対して自動ロールバックが行われるようにします。

## AC.9.7 本番環境の強化

**#AC.9.7.1 レベル: 2 役割: D/V**

本番環境がSSHアクセスをブロックし、デバッグエンドポイントを無効化し、組織の事前通知要件を伴う変更要求を必須とし(緊急時を除く)、これを確認してください。

## AC.9.8 リリースプロモーションゲート

**#AC.9.8.1 レベル: 2 役割: D/V**

プロモーションゲートには自動化されたセキュリティテスト(SAST、DAST、コンテナスキャン)が含まれ、承認にはクリティカルな問題がゼロであることを確認してください。

## AC.9.9 ワークロード、容量、およびコスト監視

**#AC.9.9.1 レベル: 1 役割: D/V**

GPU/TPUの使用率が監視され、組織で定義された閾値でアラートが発報され、容量管理ポリシーに基づいて自動スケーリングまたは負荷分散が起動されることを確認してください。

**#AC.9.9.2 レベル: 1 役割: D/V**

AIワークロードのメトリクス(推論レイテンシ、スループット、エラー率)が組織の監視要件に従って収集され、インフラストラクチャの利用状況と相關関係があることを検証する。

**#AC.9.9.3 レベル: 2 役割: V**

コストモニタリングが、組織の予算閾値に基づくアラートおよび予算超過に対する自動制御と共に、ワークロード/テナントごとの支出を追跡していることを確認してください。

#### #AC.9.9.4 レベル: 3 役割: V

容量計画が組織で定義された予測期間に基づく過去のデータを使用し、需要パターンに基づく自動リソース調達を行っていることを確認してください。

### AC.9.10 承認と監査証跡

#### #AC.9.10.1 レベル: 1 役割: D/V

環境の昇格には、組織的に定義された承認権限者による暗号署名および不变の監査証跡が必要であることを検証してください。

### AC.9.11 IaC ガバナンス

#### #AC.9.11.1 レベル: 2 役割: D/V

インフラストラクチャー・アズ・コードの変更は、メインブランチへのマージ前にピアレビュー、ならびに自動テストとセキュリティスキヤンを必須とすることを確認してください。

### AC.9.12 非本番環境におけるデータ処理

#### #AC.9.12.1 レベル: 2 役割: D/V

非本番データが組織のプライバシー要件に従って匿名化されていること、合成データ生成が行われていること、または個人識別情報(PII)が削除された完全なデータマスキングが検証されていることを確認してください。

### AC.9.13 バックアップと災害復旧

#### #AC.9.13.1 レベル: 1 役割: D/V

インフラストラクチャの構成が、3-2-1バックアップ戦略の実装に従い、組織のバックアップスケジュールに沿つて地理的に分散した地域にバックアップされていることを検証する。

#### #AC.9.13.2 レベル: 2 役割: V

組織のスケジュールに従って自動テストを通じてリカバリ手順がテストおよび検証されていることを確認し、RTOおよびRPOの目標が組織の要件を満たしていることを検証します。

#### #AC.9.13.3 レベル: 3 役割: V

災害復旧に、モデル重みの復元、GPUクラスタの再構築、およびサービス依存関係のマッピングを含むAI特有の実行手順書が含まれていることを確認してください。

### AC.9.14 コンプライアンス & ドキュメンテーション

#### #AC.9.14.1 レベル: 2 役割: D/V

組織のスケジュールに従って、SOC 2、ISO 27001、または FedRAMP の管理策に対してインフラストラクチャのコンプライアンスが評価され、自動化された証拠収集が行われていることを検証します。

#### #AC.9.14.2 レベル: 2 役割: V

インフラストラクチャのドキュメントに、組織の変更管理要件に従って更新されたネットワーク図、データフローマップ、および脅威モデルが含まれていることを確認してください。

#### #AC.9.14.3 レベル: 3 役割: D/V

インフラストラクチャの変更が、自動化されたコンプライアンス影響評価と高リスク変更に対する規制承認ワー

クロスを経ることを確認してください。

### AC.9.15 ハードウェアおよびサプライチェーン

#### #AC.9.15.1 レベル: 2 役割: D/V

AIアクセラレーターフームウェア(GPU BIOS、TPUファームウェア)が暗号署名で検証されており、組織のパッチ管理のタイムラインに従って更新されていることを確認してください。

#### #AC.9.15.2 レベル: 3 役割: V

AIハードウェアのサプライチェーンに、製造者証明書による出所確認と改ざん防止包装の検証が含まれていることを検証してください。

### AC.9.16 クラウド戦略とポータビリティ

#### #AC.9.16.1 レベル: 3 役割: V

クラウドベンダーロックイン防止には、ポータブルなインフラストラクチャ・アズ・コード、標準化されたAPI、およびフォーマット変換ツールを備えたデータエクスポート機能が含まれていることを確認してください。

#### #AC.9.16.2 レベル: 3 役割: V

マルチクラウドのコスト最適化には、リソースの無秩序な増加を防止するセキュリティコントロールと、無許可のクロスクラウド間データ転送料金を防ぐ対策が含まれていることを確認してください。

### AC.9.17 GitOps と自己修復

#### #AC.9.17.1 レベル: 2 役割: D/V

GitOpsリポジトリが、GPGキーによる署名付きコミットを必要とし、mainブランチへの直接プッシュを防止するブランチ保護ルールを設定していることを確認してください。

#### #AC.9.17.2 レベル: 3 役割: V

セルフヒーリングインフラストラクチャには、セキュリティイベントの相関分析、自動化されたインシデント対応、およびステークホルダー通知ワークフローが含まれていることを確認してください。

### AC.9.18 ゼロトラスト、エージェント、プロビジョニングおよびレジデンシー証明

#### #AC.9.18.1 レベル: 2 役割: D/V

クラウドリソースへのアクセスが、継続的な認証を伴うゼロトラスト検証を含んでいることを確認してください。

#### #AC.9.18.2 レベル: 2 役割: D/V

自動化されたインフラストラクチャプロビジョニングが、セキュリティポリシーの検証を含み、準拠していない構成に対しては展開をブロックすることを確認してください。

#### #AC.9.18.3 レベル: 2 役割: D/V

自動化されたインフラストラクチャプロビジョニングがCI/CDの過程でセキュリティポリシーを検証し、非準拠の設定がデプロイからブロックされることを確認してください。

#### #AC.9.18.4 レベル: 3 役割: D/V

保存場所の暗号学的証明によってデータ居住要件が適用されていることを検証します。

#### #AC.9.18.5 レベル: 3 役割: D/V

クラウドプロバイダーのセキュリティ評価に、エージェント固有の脅威モデリングとリスク評価が含まれていることを確認してください。

### AC.9.19 アクセス制御とアイデンティティ

**#5.1.3 レベル: 2 役割: D**

新しいプリンシパルが、本番システムアクセスを受ける前に、NIST 800-63-3 IAL-2 または同等の基準に準拠した本人確認手続きを受けることを確認してください。

**#5.1.4 レベル: 2 役割: V**

アクセスレビューが四半期ごとに実施されていることを確認し、休眠アカウントの自動検出、資格情報のローテーション強制、およびプロジェクト解除のワークフローが適用されていることを検証してください。

**#5.2.2 レベル: 1 役割: D/V**

サービスアカウントについては、最小権限の原則がデフォルトで適用され、読み取り専用の権限から開始し、書き込みアクセスには文書化された業務上の正当性が必要であることを確認してください。

**#5.3.3 レベル: 2 役割: D**

ポリシー定義がバージョン管理され、ピアレビューされ、CI/CDパイプライン内の自動テストを通じて検証されてから本番展開されることを確認してください。

**#5.3.4 レベル: 2 役割: V**

ポリシー評価の結果に決定根拠が含まれていることを確認し、それらが相関分析およびコンプライアンス報告のためにSIEMシステムに送信されることを検証します。

**#5.4.4 レベル: 2 役割: V**

ポリシー評価の遅延が継続的に監視されており、認可回避を可能にするタイムアウト条件に対して自動アラートが設定されていることを確認してください。

**#5.5.4 レベル: 2 役割: V**

検閲アルゴリズムが決定的であり、バージョン管理されていて、コンプライアンス調査およびフォレンジック分析をサポートするために監査ログを維持していることを確認してください。

**#5.5.5 レベル: 3 役割: V**

高リスクな修正イベントが、データ露出を伴わずに元のコンテンツの暗号学的ハッシュを含む適応ログを生成して、フォレンジックリトリーバルを可能にすることを検証してください。

**#5.7.5 レベル: 3 役割: V**

エージェントのエラー条件および例外処理に、インシデント分析とフォレンジック調査をサポートするための機能範囲情報が含まれていることを確認してください。

**#5.4.2 レベル: 1 役割: D/V**

モデルの出力における引用、参考文献、および情報源の帰属が呼び出し元の権限に照らして検証され、不正アクセスが検出された場合は削除されることを確認してください。

## 上に統合される新しいアイテム

**#2.3.3 レベル: 2 役割: D/V**

許可された文字セットが定期的に見直され、ビジネス要件に合致し続けていることを確認してください。

# 付録D: AI支援によるセキュアコーディングのガバナンスと検証

## 目的

この章では、ソフトウェア開発中にAI支援コーディングツールを安全かつ効果的に使用するための基準となる組織的統制を定義し、SDLC全体でのセキュリティとトレーサビリティを確保します。

## AD.1 AI支援セキュアコーディングワークフロー

既存のセキュリティゲートを弱体化させることなく、組織のセキュアソフトウェア開発ライフサイクル (SSDLC)にAIツールを統合する。

### #AD.1.1 レベル: 1 役割: D/V

文書化されたワークフローが、AIツールがコードを生成、リファクタリング、またはレビューするタイミングと方法を説明していることを確認してください。

### #AD.1.2 レベル: 2 役割: D

ワークフローが各SSDLCフェーズ(設計、実装、コードレビュー、テスト、デプロイメント)に対応していることを検証してください。

### #AD.1.3 レベル: 3 役割: D/V

AI生成コードに対して指標(例:脆弱性密度、検出までの平均時間)が収集され、人間のみのベースラインと比較されていることを確認してください。

## AD.2 AIツール資格認定と脅威モデリング

AIコーディングツールは導入前に、セキュリティ機能、リスク、およびサプライチェーンへの影響について評価を行うことを確実にしてください。

### #AD.2.1 レベル: 1 役割: D/V

各AIツールの脅威モデルが、誤用、モデル逆転、データ漏洩、および依存関係チェーンのリスクを特定していることを確認してください。

### #AD.2.2 レベル: 2 役割: D

ツールの評価に、ローカルコンポーネントの静的/動的解析およびSaaSエンドポイント(TLS、認証/認可、ログ記録)の評価が含まれていることを確認してください。

### #AD.2.3 レベル: 3 役割: D/V

評価が認識されたフレームワークに従って行われていることを確認し、主要なバージョン変更後に再実施されていることを検証してください。

## AD.3 セキュアなプロンプトおよびコンテキスト管理

AIモデルのプロンプトやコンテキストを作成する際に、秘密情報、専有コード、および個人データの漏洩を防止してください。

#### #AD.3.1 レベル: 1 役割: D/V

書面によるガイダンスが、プロンプト内で秘密情報、認証情報、または機密データの送信を禁止していることを確認してください。

#### #AD.3.2 レベル: 2 役割: D

技術的な制御(クライアント側の編集、承認されたコンテキストフィルター)が機密性の高い情報を自動的に除去することを検証してください。

#### #AD.3.3 レベル: 3 役割: D/V

プロンプトと応答がトークン化され、転送中および保存時に暗号化されていること、そして保持期間がデータ分類ポリシーに準拠していることを確認してください。

## AD.4 AI生成コードの検証

コードがマージまたはデプロイされる前に、AI出力によって導入された脆弱性を検出し修正します。

#### #AD.4.1 レベル: 1 役割: D/V

AI生成コードは常に人間によるコードレビューを受けることを確認してください。

#### #AD.4.2 レベル: 2 役割: D

AI生成コードを含むすべてのプルリクエストに対して、自動スキャナー(SAST/IAST/DAST)が実行され、重大な問題が検出された場合はマージをブロックすることを確認してください。

#### #AD.4.3 レベル: 3 役割: D/V

差分ファズテストやプロパティベースのテストによって、セキュリティに重要な動作(例: 入力検証、認可ロジック)が証明されていることを確認してください。

## AD.5 コード提案の説明可能性と追跡可能性

監査人や開発者に対して、なぜ提案がなされたのか、またそれがどのように進化したのかについての洞察を提供します。

#### #AD.5.1 レベル: 1 役割: D/V

プロンプトとレスポンスのペアがコミットIDと共にログに記録されていることを確認してください。

#### #AD.5.2 レベル: 2 役割: D

開発者が提案をサポートするモデルの引用(トレーニングの抜粋、ドキュメント)を表示できることを確認してください。

#### #AD.5.3 レベル: 3 役割: D/V

説明可能性レポートが設計成果物と共に保存され、セキュリティレビューで参照され、ISO/IEC 42001のトレーサビリティ原則を満たしていることを確認してください。

## AD.6 繼続的フィードバックとモデル微調整

モデルのセキュリティ性能を時間とともに向上させつつ、ネガティブドリフトを防止する。

### #AD.6.1 レベル: 1 役割: D/V

開発者が安全でない、または準拠していない提案にフラグを立てられること、そしてそのフラグが追跡されていることを確認してください。

### #AD.6.2 レベル: 2 役割: D

集約されたフィードバックが、査定されたセキュアコーディングコーパス(例:OWASP チートシート)を用いた定期的なファインチューニングまたはリトレーバル強化生成に役立っていることを確認してください。

### #AD.6.3 レベル: 3 役割: D/V

閉ループ評価ハーネスが、ファインチューニングのたびに回帰テストを実行することを確認してください。セキュリティ指標は、展開前に以前のベースラインを満たすかそれを超えている必要があります。

## 参考文献

- NIST AI Risk Management Framework 1.0
- ISO/IEC 42001:2023 – AI Management Systems Requirements
- OWASP Secure Coding Practices – Quick Reference Guide

## 付録E:ツールとフレームワークの例

### 目的

この章では、特定のAISVS要件の実装または達成を支援するツールおよびフレームワークの例を示します。これらは、AISVSチームやOWASP GenAIセキュリティプロジェクトによる推奨または支持と見なされるものではありません。

### AE.1 トレーニングデータガバナンスとバイアスマネジメント

データ分析、ガバナンス、およびバイアス管理に使用されるツール。

#### #AE.1.1 セクション: 1.1

データインベントリツール: データインベントリ管理ツールのようなもの...

#### #AE.1.2 セクション: 1.2

転送中の暗号化 HTTPSベースのアプリケーションにはTLSを使用し、openSSLやpythonのようなツールを利用します`ssl`ライブラリ。

### AE.2 ユーザー入力の検証

ユーザー入力を処理および検証するためのツール。

#### #AE.2.1 セクション: 2.1

プロンプトインジェクション防御ツール: NVIDIAのNeMoやGuardrails AIのようなガードレールツールを使用します。

## 付録 B: 戰略的コントロール

### C4.15 量子耐性インフラストラクチャセキュリティ

量子耐性暗号技術と量子安全プロトコルを通じて、量子コンピューティングの脅威に備えたAIインフラストラクチャを準備する。

#### #4.15.1 レベル: 3 役割: D/V

AIインフラストラクチャが、鍵交換およびデジタル署名のためにNIST承認のポスト量子暗号アルゴリズム(CRYSTALS-Kyber, CRYSTALS-Dilithium, SPHINCS+)を実装していることを確認してください。

#### #4.15.2 レベル: 3 役割: D/V

量子安全な鍵管理プロトコルを用いて、高セキュリティのAI通信のために量子鍵配送(QKD)システムが実装されていることを検証する。

#### #4.15.3 レベル: 3 役割: D/V

暗号のアジャリティフレームワークが、自動化された証明書および鍵のローテーションを伴う新しいポスト量子アルゴリズムへの迅速な移行を可能にすることを検証してください。

#### #4.15.4 レベル: 3 役割: V

量子脅威モデリングが、ドキュメント化された移行スケジュールとリスク評価を伴って、量子攻撃に対するAIインフラの脆弱性を評価していることを確認してください。

#### #4.15.5 レベル: 3 役割: D/V

ハイブリッド古典-量子暗号システムが、量子移行期間中に防御の多層化を提供し、パフォーマンス監視を行っていることを検証する。

### C4.17 ゼロ知識インフラストラクチャ

機密情報を開示することなくプライバシー保護されたAI検証および認証のためのゼロ知識証明システムを実装する。

#### #4.17.1 レベル: 3 役割: D/V

ゼロ知識証明(ZK-SNARKs)が、モデルの重みや訓練データを公開することなく、AIモデルの整合性と訓練の起源を検証することを確認します。

#### #4.17.2 レベル: 3 役割: D/V

ZKベースの認証システムが、身元に関連する情報を公開することなく、AIサービス向けにプライバシーを保護したユーザー認証を可能にすることを検証してください。

#### #4.17.3 レベル: 3 役割: D/V

プライベートセットインターフェクション(PSI)プロトコルが、個々のデータセットを公開することなく、フェデレーテッドAIにおける安全なデータマッチングを可能にすることを検証してください。

#### #4.17.4 レベル: 3 役割: D/V

ゼロ知識機械学習(ZKML)システムが、正しい計算の暗号的証明によって検証可能なAI推論を実現することを検証する。

#### #4.17.5 レベル: 3 役割: D/V

ZK-rollupsは、バッチ検証と計算負荷の軽減により、スケーラブルでプライバシー保護されたAIトランザクション処理を提供することを検証してください。

## C4.18 サイドチャネル攻撃防止

機密情報を漏洩させる可能性のあるタイミング攻撃、電力攻撃、電磁攻撃、およびキャッシュベースのサイドチャネル攻撃からAIインフラストラクチャを保護します。

### #4.18.1 レベル: 3 役割: D/V

AI推論のタイミングが定数時間アルゴリズムとパディングを使用して正規化され、タイミングに基づくモデル抽出攻撃を防止していることを検証してください。

### #4.18.2 レベル: 3 役割: D/V

パワー解析防御には、ノイズ注入、電源ラインフィルタリング、およびランダム化された実行パターンが含まれていることを確認してください。

### #4.18.3 レベル: 3 役割: D/V

キャッシュベースのサイドチャネル対策は、情報漏洩を防ぐためにキャッシュ分割、ランダム化、およびフラッシュ命令を使用していることを確認してください。

### #4.18.4 レベル: 3 役割: D/V

電磁波放射保護には、TEMPESTスタイルの攻撃を防ぐために、シールド、信号フィルタリング、およびランダム化処理が含まれていることを確認してください。

### #4.18.5 レベル: 3 役割: D/V

マイクロアーキテクチャのサイドチャネル防御には、投機的実行制御とメモリアクセスパターンの難読化が含まれていることを確認してください。

## C4.19 ニューロモルフィックおよび特殊AIハードウェアのセキュリティ

ニューロモルフィックチップ、FPGA、カスタムASIC、および光学計算システムを含む新興AIハードウェアアーキテクチャのセキュリティを確保する。

### #4.19.1 レベル: 3 役割: D/V

ニューロモルフィックチップのセキュリティには、スパイクパターンの暗号化、シナプス重みの保護、およびハードウェアベースの学習ルールの検証が含まれていることを確認してください。

### #4.19.2 レベル: 3 役割: D/V

FPGAベースのAIアクセラレータが、ビットストリーム暗号化、改ざん防止機構、および認証されたアップデートによる安全な構成読み込みを実装していることを検証してください。

### #4.19.3 レベル: 3 役割: D/V

カスタムASICのセキュリティには、オンチップセキュリティプロセッサ、ハードウェアルートオブトラスト、および改ざん検出機能を備えた安全なキー保管が含まれていることを確認してください。

### #4.19.4 レベル: 3 役割: D/V

光学計算システムが量子安全な光学暗号化、セキュアなフォトニックスイッチング、および保護された光信号処理を実装していることを検証してください。

### #4.19.5 レベル: 3 役割: D/V

ハイブリッドアナログ・デジタルAIチップが、安全なアナログ計算、保護された重みのストレージ、認証されたアナログからデジタルへの変換を含んでいることを検証してください。

## C4.20 プライバシー保護計算基盤

AI処理および分析中の機密データを保護するために、プライバシー保護計算のためのインフラストラクチャ制御を実装する。

### #4.20.1 レベル: 3 役割: D/V

同型暗号インフラストラクチャが、暗号的整合性検証および性能監視を伴う機密性の高いAIワークフロー上で暗号化計算を可能にすることを検証してください。

### #4.20.2 レベル: 3 役割: D/V

プライベート情報検索システムがアクセスパターンの暗号化による保護を伴い、クエリパターンを明らかにせずにデータベースクエリを可能にすることを検証してください。

### #4.20.3 レベル: 3 役割: D/V

安全なマルチパーティ計算プロトコルが、個々の入力や中間計算を公開することなくプライバシーを保護するAI推論を可能にすることを検証してください。

### #4.20.4 レベル: 3 役割: D/V

プライバシー保護型の鍵管理には、分散鍵生成、しきい値暗号技術、ハードウェア支援保護による安全な鍵ローテーションが含まれていることを確認してください。

### #4.20.5 レベル: 3 役割: D/V

暗号学的なセキュリティ保証を維持しながら、バッチ処理、キャッシュ、およびハードウェアアクセラレーションを通じてプライバシー保護計算の性能が最適化されていることを確認する。

### 4.9.14.9.2 1: 2 D/V: D/V

マルチクラウド展開がフェデレーテッドアイデンティティ標準(例:OIDC、SAML)を使用し、プロバイダー間で集中化されたポリシーの強制が行われていることを確認する。

### 4.9.14.9.3 1: 2 D/V: D/V

クロスクラウドおよびハイブリッドデータ転送が、カスタマー管理キーによるエンドツーエンド暗号化を使用し、管轄地域のデータ居住要件を適用していることを確認する。

### 4.9.14.9.1 1: 1 D/V: D/V

クラウドストレージ統合がエージェント制御のキー管理によるエンドツーエンド暗号化を使用していることを検証する。

### 4.9.14.9.2 1: 2 D/V: D/V

ハイブリッド展開のセキュリティ境界が明確に定義され、通信チャネルが暗号化されていることを検証します。