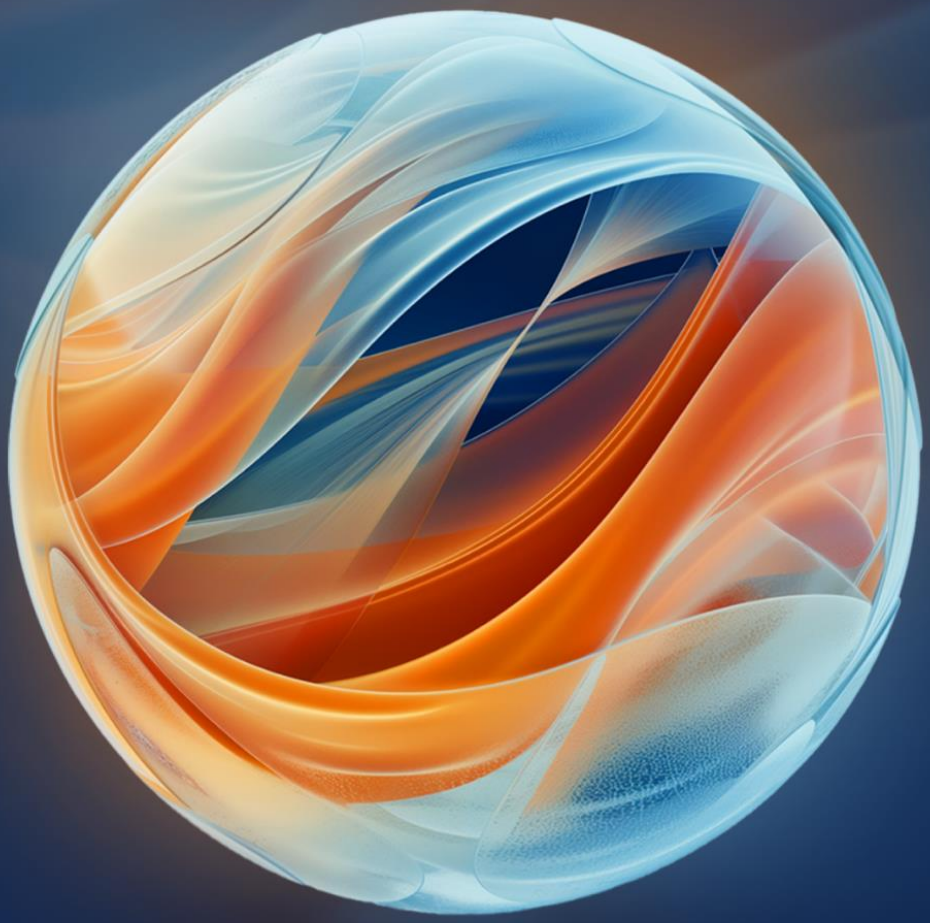
AI組織の責任：

コアセキュリティの責任



AI Organizational Responsibilities Working Group

cloud security alliance®

The permanent and official location for the AI Organizational Responsibilities Working Group is https://cloudsecurityalliance.org/research/working-groups/ai-organizational-responsibilities。

謝辞

Lead Authors

Jerry Huang
Ken Huang

Contributors/Co-Chairs

Ken Huang
Nick Hamilton
Chris Kirschke
Sean Wright

Reviewers

Candy Alexander
Ilango Allikuzhi
Eray Altili
Aakash Alurkar
Romeo Ayalin
Renu Bedi
Saurav Bhattacharya
Sergei Chaschin
Hong Chen
John Chiu
Satchit Dokras
Rajiv Gunja
Hongtao Hao, PhD
Grace Huang
Onyeka Illoh
Krystal Jackson
Arvin Jakkamreddy Reddy
Simon Johnson
Gian Kapoor
Ben Kereopa-Yorke
Chris Kirschke
Madura Malwatte
Madhavi Najana
Rajith Narasimhaiah
Gabriel Nwajiaku
Govindaraj Palanisamy
Meghana Parwate
Paresh Patel
Rangel Rodrigues
Michael Roza
Lars Ruddigkeit
Davide Scatto
Maria Schwenger Mj
Bhuvaneswari Selvadurai
Himanshu Sharma
Akshay Shetty
Nishanth Singarapu
Abhinav Singh
Dr. Chantal Spleiss
Patricia Thaine
Eric Tierling
Ashish Vashishtha
Peter Ventura
Jiewen Wang
Wickey Wang
Udith Wickramasuriya
Sounil Yu

CSA Global Staff

Marina Bregkou Sean Heide Alex Kaluza Claire Lehnert Stephen Lumpe

日本語版提供に際しての告知及び注意事項

本書「AI組織の責任：コアセキュリティの責任」は、Cloud Security Alliance (CSA)が公開している「AI Organizational Responsibilities: Core Security Responsibilities」の日本語訳です。本書は、CSAジャパンが、CSAの許可を得て翻訳し、公開するものです。原文と日本語版の内容に相違があった場合には、原文が優先されます。
翻訳に際しては、原文の意味および意図するところを、極力正確に日本語で表すことを心がけていますが、翻訳の正確性および原文への忠実性について、CSAジャパンは何らの保証をするものではありません。
この翻訳版は予告なく変更される場合があります。以下の変更履歴（日付、バージョン、変更内容）をご確認ください。

変更履歴

日付	バージョン	変更内容
2024年8月14日	日本語版1.0	初版発行

CSAジャパン成果物の提供に際しての制限事項

日本クラウドセキュリティアライアンス（CSAジャパン）は、本書の提供に際し、以下のことをお断りし、またお願いします。以下の内容に同意いただけない場合、本書の閲覧および利用をお断りします。

１．　責任の限定
CSAジャパンおよび本書の執筆・作成・講義その他による提供に関わった主体は、本書に関して、以下のことに対する責任を負いません。また、以下のことに起因するいかなる直接・間接の損害に対しても、一切の対応、是正、支払、賠償の責めを負いません。
(1)　本書の内容の真正性、正確性、無誤謬性
(2)　本書の内容が第三者の権利に抵触しもしくは権利を侵害していないこと
(3)　本書の内容に基づいて行われた判断や行為がもたらす結果
(4)　本書で引用、参照、紹介された第三者の文献等の適切性、真正性、正確性、無誤謬性および他者権利の侵害の可能性

２．　二次譲渡の制限
本書は、利用者がもっぱら自らの用のために利用するものとし、第三者へのいかなる方法による提供も、行わないものとします。他者との共有が可能な場所に本書やそのコピーを置くこと、利用者以外のものに送付・送信・提供を行うことは禁止されます。また本書を、営利・非営利を問

わず、事業活動の材料または資料として、そのまま直接利用することはお断りします。
ただし、以下の場合は本項の例外とします。
(1)　本書の一部を、著作物の利用における「引用」の形で引用すること。この場合、出典を明記してください。
(2)　本書を、企業、団体その他の組織が利用する場合は、その利用に必要な範囲内で、自組織内に限定して利用すること。
(3)　CSAジャパンの書面による許可を得て、事業活動に使用すること。この許可は、文書単位で得るものとします。
(4)　転載、再掲、複製の作成と配布等について、CSAジャパンの書面による許可・承認を得た場合。この許可・承認は、原則として文書単位で得るものとします。

３．　本書の適切な管理
(1)　本書を入手した者は、それを適切に管理し、第三者による不正アクセス、不正利用から保護するために必要かつ適切な措置を講じるものとします。
(2)　本書を入手し利用する企業、団体その他の組織は、本書の管理責任者を定め、この確認事項を順守させるものとします。また、当該責任者は、本書の電子ファイルを適切に管理し、その複製の散逸を防ぎ、指定された利用条件を遵守する（組織内の利用者に順守させることを含む）ようにしなければなりません。
(3)　本書をダウンロードした者は、CSAジャパンからの文書（電子メールを含む）による要求があった場合には、そのダウンロードしまたは複製した本書のファイルのすべてを消去し、削除し、再生や復元ができない状態にするものとします。この要求は理由によりまたは理由なく行われることがあり、この要求を受けた者は、それを拒否できないものとします。
(4)　本書を印刷した者は、CSAジャパンからの文書（電子メールを含む）による要求があった場合には、その印刷物のすべてについて、シュレッダーその他の方法により、再利用不可能な形で処分するものとします。

４．　原典がある場合の制限事項等
本書がCloud Security Alliance, Inc.の著作物等の翻訳である場合には、原典に明記された制限事項、免責事項は、英語その他の言語で表記されている場合も含め、すべてここに記載の制限事項に優先して適用されます。

５．　その他
その他、本書の利用等について本書の他の場所に記載された条件、制限事項および免責事項は、すべてここに記載の制限事項と並行して順守されるべきものとします。本書およびこの制限事項に記載のないことで、本書の利用に関して疑義が生じた場合は、CSAジャパンと利用者は誠意をもって話し合いの上、解決を図るものとします。

その他本件に関するお問合せは、info@cloudsecurityalliance.jp までお願いします。

日本語版作成に際しての謝辞

「AI組織の責任：コアセキュリティの責任」は、CSAジャパン会員の有志により行われました。作業は全て、個人の無償の貢献としての私的労力提供により行われました。なお、企業会員からの参加者の貢献には、会員企業としての貢献も与っていることを付記いたします。
以下に、翻訳に参加された方々の氏名を記します。（氏名あいうえお順・敬称略）

石井　英男
笠松　隆幸
仲上　竜太
三井　陽一　　CISSP, CCSP, CISA, CISM, CDPSE
諸角　昌宏

目次

エグゼクティブサマリー

本ホワイトペーパーは、人工知能（AI）と機械学習（ML）システムの開発と展開における組織の責任について、情報セキュリティとサイバーセキュリティの側面に焦点を当てた作業草案です。本ホワイトペーパーは、データ保護メカニズム、モデルの脆弱性管理、Machine Learning Operations（MLOps）パイプラインのセキュリティ強化、AIのトレーニングと責任ある展開のためのガバナンスポリシーなど、中核的なセキュリティ分野における専門家が推奨するベストプラクティスを統合しています。

本ホワイトペーパーで議論されている主なポイントは以下の通りです。

- **データセキュリティとプライバシー保護**：AIトレーニングにおけるデータの真正性、匿名化、仮名化、データの最小化、アクセス制御、安全な保存と送信の重要性。
- **モデルセキュリティ**：アクセス制御、セキュアなランタイム環境、脆弱性とパッチの管理、MLOpsパイプラインのセキュリティ、AIモデルガバナンス、セキュアなモデルデプロイメントなど、モデルセキュリティのさまざまな側面をカバー。
- **脆弱性管理**：脆弱性を効果的に管理するためのAI/ML資産インベントリ、継続的な脆弱性スキャン、リスクに基づく優先順位付け、修復の追跡、例外処理、報告指標の重要性について議論。

本ホワイトペーパーでは、定量化可能な評価基準、役割定義のためのRACI（Responsible, Accountable, Consulted, Informed）モデル、ハイレベルの実装戦略、継続的な監視と報告の仕組み、アクセス制御のマッピング、基礎的なガードレールの遵守を用いて、各責任を分析しています。これらは、NIST AI RMF、NIST SSDF、NIST 800-53、CSA CCMなどの業界のベストプラクティスや標準に基づいています。

本ホワイトペーパーは、セキュリティとコンプライアンスに関するこれらの主要分野にわたる推奨事項を概説することにより、企業が責任ある安全なAIの設計、開発、デプロイメントの義務を果たすための指針を示すことを目的とします。

はじめに

本ホワイトペーパーでは、人工知能（AI）および機械学習（ML）、データセキュリティ、モデルセキュリティ、脆弱性管理に関する企業の「コアセキュリティ責任」として我々が定義するものに焦点を当てています。組織には、安全で安心なAIの実践を支持する義務があるため、本ホワイトペーパーと本シリーズの他の2つのホワイトペーパーは、企業がそのような組織的責任を果たすための青写真を提供します。具体的には、本ホワイトペーパーは、データ保護メカニズム、モデルの脆弱性管理、MLOpsパイプラインのセキュリティ強化、および責任を持ってAIをトレーニングし導入するためのガバナンスポリシーといった、中核的なセキュリティ分野において専門家が推奨するベストプラクティスを統合しています。本シリーズの他の2つのホワイトペーパーでは、企業向けの安全なAIの開発とデプロイのさらなる側面について論じています。3つのホワイトペーパーでセキュリティとコンプライアンスに関するこれらの重要な分野にわたる推奨事項を概説することで、本シリーズは、企業が責任ある安全なAIの設計、開発、デプロイの義務を果たすための指針を示すことを目的としています。

AI 責任共有モデル

AI責任共有モデルは、サービスモデル（SaaS、PaaS、IaaS）によって異なりますが、AIプラットフォームプロバイダ、AIアプリケーションオーナー、AI開発者、AI利用者間のタスク分担の概要を示しています。

AIアプリケーションの安全な運用には、複数の利害関係者の協力が必要です。AIの文脈では、AIサービスの利用者、AIアプリケーションの所有者と開発者、AIプラットフォームプロバイダという3つの主要な関係者の間で責任が分担されます。

AI対応への統合を評価する際には、責任共有モデルを理解し、各当事者が担当する具体的なタスクを明確にすることが極めて重要です。

AI対応アプリケーションの主要レイヤー

1. **AI**プラットフォーム：
    - このレイヤーは、アプリケーションにAI機能を提供します。AIモデル、トレーニングデータ、コンフィギュレーション設定をホストするインフラの構築と保護が含まれます。
    - セキュリティに関する考慮事項には、AIモデルによって生成される悪意のある入力や出力からの保護が含まれます。AIの安全機構は、憎悪や脱獄などの潜在的な有害入出力から保護する必要があります。
    - AIプラットフォームレイヤーには以下のタスクがあります。
        - モデルの安全性とセキュリティ
        - モデルのチューニング
        - モデル説明責任
        - モデルの設計と実装
        - モデルトレーニングとガバナンス
        - AIコンピュートおよびデータインフラ

2. **AI**アプリケーションレイヤー：
    - AIアプリケーションレイヤーは、AIの能力を活用してユーザーとインターフェースを取ります。その複雑さは多岐にわたります。最も基本的なレベルでは、スタンドアロンのAIアプリケーションは、APIの集合体への導管として機能し、ユーザーからのテキストプロンプトを処理し、応答のために基礎となるモデルにリレーします。より洗練されたAIアプリケーションは、永続化レイヤー、セマンティック・インデックス、より広範なデータソースへのアクセスを提供するプラグインなどの要素を利用して、これらのプロンプトに追加のコンテキストを付加することができます。最先端のAIアプリケーションは、既存のアプリケーションやシステムとシームレスに統合できるように設計されており、テキスト、オーディオ、ビジュアルの入力を包含して多様なコンテンツ出力を生成するマルチモーダルアプローチを可能にします。
    - AIアプリケーションのオーナーとして、シームレスなユーザー体験を保証し、あらゆる機能やサービスの追加に対応します。AIアプリケーションを有害な活動から守るためには、堅牢なアプリケーション安全システムを確立することが不可欠です。生成AI（GenAI）システムは、AIモデルにディスパッチされるプロンプトで使用されるコンテンツを徹底的に検査する必要があります。さらに、プラグインや関数のようなアドオン、データコネクター、他のAIアプリケーションとの相互作用（*AIオーケストレーション*と呼ばれるプロセス）とのやり取りを精査しなければなりません。IaaS（Infrastructure-as-a-Service）またはPaaS（Platform-as-a-Service）サービス上でAIアプリケーションを開発する場合、専用のAIコンテンツ安全機能を統合することが望ましいです。特定の要件に応じて、保護を強化するための追加機能を実装することもできます。
    - AIアプリケーションには以下のタスクがあります：
        - AIプラグインとデータ接続
        - アプリケーションの設計と実装
        - アプリケーションインフラ
        - AI安全システム
3. **AI**利用レイヤー：
    - AI利用レイヤーはAI機能の応用と消費の概要を説明します。生成AIは、API、コマンドプロンプト、GUIのような従来のインターフェースとは異なる、革新的なユーザーとコンピュータの対話モデルを導入しています。この新しいインターフェースはインタラクティブで適応性があり、コンピュータの能力をユーザーの意図に合わせて成形します。ユーザーがシステムのデザインや機能に合わせることを要求するような以前のインターフェースとは異なり、生成AIインターフェースはユーザーとの対話を優先します。これにより、ユーザーの入力がシステムの出力を大幅に形成することが可能になり、個人、データ、企業資源を保護する安全機構の重要性が強調されます。
    - AIの利用におけるセキュリティへの考慮事項は、あらゆるコンピューターシステムと同様であり、IDアクセス管理、デバイスセキュリティ、モニタリング、データガバナンス、管理統制のための強固な対策に依存します。
    - ユーザーの行動がシステムのアウトプットに与える影響が大きいことを考えると、ユーザーの行動と責任にもっと焦点を当てる必要があります。許容される使用に関するポリシーを改訂し、従来のITアプリケーションとAIによって強化されたアプリケーションの区別についてユーザーに知らせることが不可欠です。この教育では、セキュリティ、プライバシー、倫理基準に関するAI特有の問題を取り上げるべきです。さらに、欺くために巧妙に捏造されたテキスト、音声、映像、その他のメディアを含む可能性のあるAI主導型攻撃

について、ユーザーの意識を高めることも重要です。
    - AI利用レイヤーには以下のタスクがあります：
        - ユーザートレーニングと説明責任
        - 許容される使用に関するポリシーと管理者による管理
        - アイデンティティアクセス管理（IAM）とデバイス制御
        - データガバナンス

この責任分担モデルは、役割分担を明確にし、職務の明確な分離を保証することで、AI技術の安全かつ効果的な利用に貢献することを忘れてはなりません。作業負荷の責任分担は、サービスモデルに基づくAI統合のタイプによって異なります。

1. **Software as a Service（SaaS）**：
    - SaaSベースのAI統合では、AIプラットフォームプロバイダが基盤となるインフラ、セキュリティ管理、コンプライアンス対策の管理責任を負います。
    - ユーザーとしては、特定の要件に合わせてAIアプリケーションを設定し、カスタマイズすることに主眼が置かれます。
2. **Platform as a Service（PaaS）**：
    - PaaSベースのAIプラットフォームは、その中間を提供します。プロバイダーが中核となるAI機能を管理する一方で、構成やカスタマイズはある程度コントロールできます。
    - AIモデルの安全な使用を保証し、トレーニングデータを扱い、モデルの動作（重みやバイアスなど）を調整する責任があります。
3. **Infrastructure as a Service（IaaS）**：
    - IaaSシナリオでは、インフラをよりコントロールできます。しかし、これはより多くの管理責任を負うことを意味します。
    - AIモデル、トレーニングデータ、インフラのセキュリティなど、スタック全体を管理します。

データ中心のAIシステムの基本構成要素

データ中心のAIシステムの基本構成要素は、データとモデル管理のライフサイクル全体を網羅します。これらのコンポーネントが連携することで、データを処理し、価値ある洞察や自動化された意思決定を提供できる、安全で効果的なAIシステムが構築されます。

- **生データ**：様々な情報源から収集された未加工の初期データ。
- **データの準備**：生データをクリーニングし、構造化された形式に整理するプロセス。
- **データセット**：分析およびモデルトレーニングのために用意されたデータのコレクション。
- **データとAIのガバナンス**：データの品質と倫理的なAIの使用を保証するための方針と手順。
- **機械学習アルゴリズム**：データを解釈するために使用される計算手法。
- **評価**：機械学習モデルの性能を評価。
- **機械学習モデル**：データセットで学習したアルゴリズムの出力。
- **モデル管理**：機械学習モデルのライフサイクル管理。
- **モデルのデプロイメントと推論**：予測や決定を行うためのモデルの実装。
- **推論の結果**：展開されたモデルによって生成された結果。
- **機械学習オペレーション（MLOps）**：AIモデルのデプロイと保守のためのプラクティス。
- **データとAIプラットフォームのセキュリティ**：脅威からシステムを保護するための対策。

データオペレーション：データの取得と変換に加え、データのセキュリティとガバナンスを保証します。MLモデルの有効性は、データパイプラインの完全性と強化されたDataOpsフレームワークにかかっています。

モデルの運用：予測MLモデルの作成、モデルマーケットプレイスからの調達、またはOpenAIやFoundation Model APIを通じて提供されるような大規模言語モデル（LLM）の利用が含まれます。モデル開発は反復プロセスであり、様々な実験条件を文書化し評価する体系的なアプローチが必要です。

モデルのデプロイとサービング：モデルコンテナの安全な構築、分離され保護されたモデルのデプロイ、自動スケーリング、レート制限、アクティブモデルの監視の実装を含みます。また、RAG（Retrieval Augmented Generation）アプリケーションにおける高可用性、低レイテンシーのサービスのための機能と特徴の提供や、モデルをプラットフォームの外部に展開したり、カタログからのデータ機能を必要としたりする他のアプリケーションに必要な機能の提供も含まれます。

運用とプラットフォーム：プラットフォームの脆弱性、アップデート、モデルの分離、およびシステムコントロールの管理、ならびに安全なアーキテクチャフレームワーク内での許可されたモデルアクセスの実施をカバーします。さらに、継続的インテグレーション/継続的デプロイメント（CI/CD）のための運用ツールのデプロイメントを含み、セキュアなML運用（MLOps）のために、ライフサイクル全体が、開発、ステージング、本番という別々の実行環境にわたって確立された標準に準拠していることを保証します。

表1は、データ中心AIシステムの中核となる側面とオペレーションを整合させ、それぞれの役割と相互依存関係を強調したものです。

基礎コンポーネント	説明
データ運用	データの取り込み、変換、セキュリティ、およびガバナンス。
モデル運用	MLモデルの構築、取得、実験。
モデルの展開とサービング	MLモデルのセキュアなデプロイメント、サービング、モニタリング。
オペレーションとプラットフォーム	プラットフォームのセキュリティ、モデルの分離、MLOpsのためのCI/CD。

表1：データ中心のAIシステム・コンポーネントと相互接続された役割のマッピング

表2は、AI/MLシステムの各段階における潜在的なセキュリティリスクと脅威を、これらの懸念に対処するための例と推奨される緩和策とともにまとめたものです。

システムステージ	システムコンポーネント	潜在的なセキュリティリスク	脅威	緩和策
データオペレーション	生データ、データプレップ、データセット	データの損失：データの不正な削除または破損。データ汚染：故意にデータを操作し、モデルの完全性を損なうこと。 コンプライアンスの課題：データ保護に関する規制要件を満たすことができない。	データの侵害／汚染：攻撃者は偽のデータを注入したり、既存のデータを改ざんしたりする可能性がある。	堅牢なデータガバナンスフレームワークの導入 異常検知システムの導入 復旧プロトコルを確立し、定期的にデータのバックアップを行う。
モデルオペレーション	MLアルゴリズム、モデル管理	モデルの窃盗：プロプライアティ・モデルを盗むこと。 不正アクセス：許可なくモデルにアクセスすること。	APIアクセスによる攻撃：APIの脆弱性を悪用してモデルにアクセスしたり、操作したりする。 モデルの盗用（抽出）：モデルを複製して不正に使用すること。	アクセス制御と認証メカニズムの強化：APIエンドポイントの暗号化とレート制限による安全性の確保。システムの定期的なアップデートとパッチ適用
モデルのデプロイメントとサービング	モデルサービング、推論応答	不正アクセス：モデルサービングインフラへの無許可のアクセス。データ漏洩：設定ミスによる機密情報の漏洩。	モデルを騙すこと（evasion）：モデルから特定の出力を得るために入力を変更すること。 トレーニングデータの復元（inversion）：モデルから私的なトレーニングデータを抽出すること。	コンテナ化およびネットワークセグメンテーションを含む、安全なデプロイの実践。モデルの相互作用を積極的に監視し、ログに記録する。レート制限と異常検知の実装。

オペレーションとプラットフォーム	MLオペレーション、データおよびAIプラットフォームのセキュリティ	不十分な脆弱性管理：既知の脆弱性にタイムリーに対処していない。 モデルの分離の問題：モデルを適切に分離できず、相互汚染の可能性がある。	MLのサプライチェーン攻撃：サードパーティのコンポーネントに脆弱性やバックドアを導入する。 モデル汚染（ポイズニング）：学習データを破壊し、誤分類やシステム停止を引き起こす。	継続的な脆弱性管理とパッチ適用。 一貫したデプロイのためのCI/CDプロセス分離制御と安全なアーキテクチャ設計

表2：AI/MLセキュリティリスクの概要

各責任を以下の側面から分析します。

1. 評価基準： AIの責任について議論する際には、AIシステムのセキュリティ上の影響を評価するための定量化可能な指標を検討します。これらの側面を定量化することで、利害関係者はAI技術の関連リスクとそのリスクへの対処方法について理解を深めることができます。組織は、セキュリティと信頼性を確保するために、AIシステムを頻繁に評価しなければなりません。システムが攻撃をどの程度処理できるか（敵対的堅牢性）、機密データを漏えいしないか、ミスを犯す頻度（誤検出率）、学習データが信頼できるか（データの完全性）など、測定可能な事柄を評価する必要があります。組織のセキュリティ計画の一環として、これらの重要な対策を評価・監視することは、AIシステムの全体的なセキュリティ態勢の改善に役立ちます。

2. RACIモデル：この指標は、AIの意思決定と監督に関して、実行/執行責任（Responsible）、説明責任（Accountable）、相談先（Consulted）、報告先（Informed）の、どの責任を誰が担うかを明確にするのに役立ちます。RACIモデルを適用することで、AIガバナンスにおける役割と責任が明確になります。この責任分担は、安全なAIシステムにとって不可欠です。組織の規模や事業内容によっては、本ホワイトペーパーに記載されている具体的な役割やチームはあくまで参考であることを理解することが重要です。まず、重要な責任を明確に概説することに重点を置くべきです。そして、その責任に対応する適切な役割と、その役割を担うチームを決定します。チーム間で重複する責任もあるかもしれません。ここで定義するRACIフレームワークは、組織が独自のRACIモデルを開発する際の助けとなるよう、最初の役割とチームの役職名を提供することを目的としています。しかし、その実施方法は、各企業の組織構造や優先事項によって異なる可能性があります。

3. ハイレベルの実装戦略：この指標は、サイバーセキュリティへの配慮をソフトウェア開発ライフサイクル（SDLC）にシームレスに統合するための戦略を概説します。組織は、データとシステムの機密性、完全性、可用性を確保する CIA 原則の実施を優先しなければなりません。ユーザー権限を管理し、不正アクセスを防止するために、アクセス制御の仕組みを厳格に実装しなければなりません。堅牢な監査機構は、システムの活動を追跡し、疑わしい行動を迅速に検出しなければなりません。影響評価では、AI システムの機密情報を保護するために、脆弱性の特定と脅威の軽減に重点を置いて、潜在的なサイバーセキュリティリスクを評価する必要があります。

4. 継続的なモニタリングとレポーティング：この指標は、継続的なモニタリングとレポーティングは、AIシステムの継続的なセキュリティ、安全性、パフォーマンスを保証します。重要なコンポーネントには、リアルタイムのモニタリング、モデルパフォーマンスの低下やセキュリティインシデントに対するアラート、監査証跡/ログ、定期的なレポーティングが含まれます。継続的なモニタリングとレポーティングは、組織が透明性を維持し、パフォーマンスと説明責任を強化し、AIシステムに対する信頼を構築するのに役立ちます。

5. アクセス制御：この指標は、AIシステムの安全性を確保する上で極めて重要です。これには、強力なAPI認証/認可ポリシー、モデルレジストリの管理、データリポジトリへのアクセス制御、継続的インテグレーションとデプロイメントパイプライン（CI/CD）の監督、シークレットの取り扱い、特権アクセスの管理などが含まれます。AIパイプラインのさまざまな部分についてユーザーの役割と権限を定義することで、機密データを保護し、適切な権限なしにモデルを改ざんしたり、アクセスしたりすることができなくなります。強力なアイデンティティとアクセス管理を実装することで、知的財産を保護するだけでなく、AIワークフロー全体の説明責任も確保できます。

6. 基本的なガバナンス、リスクとコンプライアンス、セキュリティ、安全性、倫理的なガードレールの遵守：
以下のような業界のベストプラクティスや規制要件に基づくガードレールの遵守を重視します：

- 安全なソフトウェア開発のためのNIST SSDF
- NIST人工知能リスクマネジメントフレームワーク（AI RMF）
- ISO/IEC 42001:2023 AIマネジメントシステム（AIMS）
- ISO/IEC 27001:2022 情報セキュリティマネジメントシステム（ISMS）
- ISO/IEC 27701:2019 プライバシー情報管理システム（PIMS）
- ISO 31700-1:2023 消費者保護 消費財・サービスのプライバシー・バイ・デザイン
- LLMアプリケーションのOWASP Top10
- NIST SP 800-53 Rev.5 情報システムおよび組織のためのセキュリティおよびプライバシー管理
- データの匿名化および仮名化に関する一般データ保護規則（GDPR）とガイダンス
- クラウドベースのサービスにおけるトークン化のガイダンス

前提条件

本ホワイトペーパーは産業中立的な立場を前提としており、特定の産業に偏ることなく、様々な分野に適用できるガイドラインや推奨事項を提供しています。

想定読者

このホワイトペーパーは、それぞれ異なる目的と関心を持つ多様な読者に対応することを意図しています。

1. 最高情報セキュリティ責任者（CISO）：このホワイトペーパーは、CISOの懸念と責任に対応するために特別に設計されています。AIシステムにセキュリティの基本原則を統合するための貴重な洞察を提供します。多くの組織でAI最高責任者（CAIO）の役割が生まれつつあり、近い将来、本ホワイトペーパーで定義されている関連責任の大部分がCISOからCAIOに移行することが予想されることにご留意ください。

2. AI研究者、エンジニア、データ専門家、サイエンティスト、アナリスト、開発者：本ホワイトペーパーは、AI研究者やエンジニアが倫理的で信頼できるAIシステムを開発するための包括的なガイドラインとベストプラクティスを提供します。責任あるAI開発を保証するための重要なリソースとなります。

3. ビジネスリーダーと意思決定者：CIO、CPO、CDO、CRO、CEO、CTOなどのビジネスリーダーや意思決定者向けに、AIシステムの開発、展開、ライフサイクル管理に関連するサイバーセキュリティ戦略に関する重要な情報と認識を提供します。

4. **政策立案者と規制当局者**：政策立案者と規制当局は、AIの倫理、安全性、制御に関する政策と規制の枠組みを形成するのに役立つ重要な洞察を提供しているため、本論文は非常に貴重です。AIガバナンスの領域において、十分な情報に基づいた意思決定を行うための指針として機能します。

5. **投資家と株主**：株主は、責任あるAIの実践に対する組織のコミットメントを示す本ホワイトペーパーを高く評価するでしょう。倫理的なAI開発を保証するためのガバナンスメカニズムが強調されており、これは投資判断に不可欠となり得ます。

6. **顧客と一般人**：本ホワイトペーパーは、安全なAIモデルを開発する際の組織の価値観や原則について、顧客や一般人に透明性を提供します。

責任と役割の定義

以下の表は、AIテクノロジーを統合または運用する組織で見られる様々な役割を示した一般的なガイドです。各組織は、独自の業務ニーズ、文化、AIイニシアチブの具体的な要求を反映して、これらの役割と関連する責任を異なった形で定義する可能性があることを認識することが不可欠です。したがって、この表は、AIのガバナンス、技術支援、開発、戦略的マネジメントにおける潜在的役割の基礎的理解を提供するものではありますが、あくまで参考目的のためのものです。各組織は、戦略目標や運用フレームワークと整合した構造と責任を確保しながら、特定の要件に最適なようにこれらの役割を適応させ、調整することが推奨されます。AI技術の進化に伴い、新たな役割を定義することも可能です。

マネジメントと戦略

役名	役割
最高データ責任者（**CDO**）	エンタープライズデータ管理、ポリシー作成、データ品質、ライフサイクルを監督する。
最高技術責任者（**CTO**）	技術戦略を主導し、技術開発を監督する。
最高情報セキュリティ責任者（**CISO**）	情報セキュリティ戦略と運用を監督する。
事業部門リーダー	事業部門を指揮し、AIイニシアチブを事業目標と整合させる。
最高**AI**責任者（**CAIO**）	組織内におけるAI技術の戦略的導入と管理を担当する。

マネジメント	CEO、COO、CIO、CTO、CISO、CAIO、CFOなどの組織目標との整合性を確保しながら、全体的な戦略を監督・指導する。
チーフクラウドオフィサー	クラウド戦略をリードし、クラウドリソースがビジネスおよび技術目標に合致していることを確認する。
チーフアーキテクト	アーキテクチャ戦略をリードし、企業の標準、プロセス、手順、目標に沿ったテクノロジーアーキテクチャを設計する。技術選択を行い、設計の品質と実装を監督し、組織内のハイパフォーマンスアーキテクトを育成する。

ガバナンスとコンプライアンス

役名	役割	カテゴリー名
データガバナンス委員会	データガバナンスと使用に関する方針と基準を設定する。	ガバナンスとコンプライアンス
データ保護責任者	データ保護戦略およびデータ保護法規の遵守を監督する。	ガバナンスとコンプライアンス
最高プライバシー責任者	個人情報保護に関する法律および規制の遵守を確保する。	ガバナンスとコンプライアンス
法務チーム／部門	AIの導入と使用に関する法的ガイダンスを提供する。法的/規制上の義務について伝える。AIベンダーとの基本契約における適切な条項を確保する。	ガバナンスとコンプライアンス
コンプライアンスチーム／部門	社内外のコンプライアンス要件を確実に遵守する。	ガバナンスとコンプライアンス
データガバナンスオフィサー	組織内のデータガバナンスを管理し、ポリシー、データプライバシー法、規制遵守要件の遵守を確保する。	ガバナンスとコンプライアンス
情報セキュリティオフィサー	ISSO、ISM、ISS を含む情報システムまたはプログラムについて、適切な運用セキュリ ティ態勢が維持されていることを保証する権限を有する承認責任者、管理当局者、または情報システム所有者。	ガバナンスとコンプライアンス

技術およびセキュリティ

役名	役割
セキュリティオペレーションチーム	データとシステムを保護するためのセキュリティプロトコルを導入し、監視する。
ネットワークセキュリティチーム	脅威や脆弱性からネットワークを保護する。
クラウドセキュリティチーム	クラウドベースのリソースとサービスのセキュリティを確保する。
サイバーセキュリティチーム	サイバー脅威、脆弱性、組織資産への不正アクセスから保護する。
ITオペレーションチーム	ITインフラストラクチャをサポートし、運用と安全性を維持する。
ネットワークセキュリティオフィサー	ネットワークのセキュリティを監督し、データ保護と脅威の緩和を確保する。
ハードウェアセキュリティチーム	物理的なハードウェアを改ざんや不正アクセスから保護する。
システム管理者	最適なパフォーマンスとセキュリティを実現するために、ITシステムとサーバーを管理・設定する。

オペレーションと開発

役名	役割
データ管理者	安全な保管、送信、データ保管、ビジネスルールの実施に責任を負う。データ所有者に代わってデータを取得、操作、保管、移動する組織または個人を指す。
AI開発チーム	AIモデルとソリューションの開発および実装。
品質保証チーム	AIアプリケーションやシステムの品質をテストし、保証する。
AIオペレーションチーム	AIシステムのパフォーマンスと信頼性を管理する。
アプリケーション開発チーム	必要に応じてAI機能を統合したアプリケーションを開発する。
AI/MLテストチーム	AI/MLモデルの精度、性能、信頼性のテストを専門とする。

開発オペレーション（**DevOps**）チーム	デプロイ効率を高め、運用の安定性を維持する。
開発セキュリティオペレーション（**DevSecOps**）チーム	ソフトウェア開発ライフサイクル（SDLC）全体にわたってセキュリティを実装する。
AIメンテナンスチーム	AIシステムとモデルがアップデートされ、最適化され、導入後に正しく機能することを確認する。
プロジェクト管理チーム	AIプロジェクトの立ち上げから完了までを監督し、目的とスケジュールを確実に達成する。
運用スタッフ	日々の業務をサポートし、AI技術の円滑な統合と機能を確保する。
データサイエンスチーム	AIモデルのトレーニングや分析に使用するデータを収集し、準備する。
コンテナ管理チーム	コンテナ化されたアプリケーションを管理し、デプロイとスケーラビリティを促進する。
ITオペレーションチーム	ITインフラの運用を保証し、AIとテクノロジーのニーズをサポートする。
AI開発マネージャー	AI開発プロジェクトをリードし、チームを成功に導く。
AIオペレーション部長	AIに関連する業務を指揮し、AIソリューションの効率性と有効性を確保する。

引用

以下の文書は、この文書を適用し理解するために不可欠なものです。

- Generative AI safety: Theories and Practices
- OpenAI Preparedness Framework
- Applying the AIS Domain of the CCM to Generative AI
- EU AI Act
- Biden Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence
- OWASP Top 10 for LLM Applications
- CSA Cloud Controls Matrix (CCM v4)
- MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems)
- NIST Secure Software Development Framework(SSDF)
- NIST Artificial Intelligence Trustworthiness and Risk Management Framework-
- General Data Protection Regulation (GDPR)
- OWASP LLM AI Cybersecurity & Governance Checklist
- OWASP Machine Learning - Top 10
- WEF Briefing Papers
- Building the AI-Powered Organization

1. AIトレーニングにデータセキュリティとプライバシーを組み込む

AIは複雑なモデル中心のアプローチからデータ中心のアプローチへと移行しつつあります。AIは現在、小規模なデータセットで学習させた複雑なモデルに頼るのではなく、大規模なデータセットやオープンなデータストリームを活用しています。しかし、このようなデータ中心のパラダイムは、データのプライバシー、セキュリティ、偏見、適切な使用に関する懸念も引き起こし、AIコミュニティは責任を持って対処しなければなりません。データはAIを変貌させつつありますが、我々はそれが倫理的に用意され、管理されることを保証しなければなりません。

以下のセクションでは、AI組織におけるトレーニングデータのセキュリティとプライバシーの確保に関連する重要なカテゴリーについて説明します。これらのカテゴリーには、「データの真正性」、「匿名化／仮名化」、「データの最小化」、「データへのアクセス制御」、「安全な保管と送信」が含まれます。各カテゴリーは、定量化可能な評価基準、RACIモデルによる責任の明確化、ハイレベルな実装戦略、継続的な監視と報告の仕組み、アクセス制御のマッピング、業界のベストプラクティスに基づく基礎的なガードレールの遵守によって徹底的に分析されます。この包括的なアプローチにより、AIの進化を促進する重要な資産を管理するために、構造化され説明責任のある、効率的なフレームワークが保証されるとともに、倫理的な義務やオペレーショナルエクセレンスにも合致します。

1.1 データの真正性と同意の管理

AIにおけるデータの真正性とは、AIモデルのトレーニング、テスト、デプロイに使用されるデータが本物であり、正確で信頼できることを保証することです。AIアルゴリズムに誤解を与えたり、不正確で偏った、あるいは信頼性の低いモデル出力をもたらすような方法で、データが改ざんされたり変更されたりしていないことを検証することといえます。

AIモデルはデータの品質と完全性に大きく依存しているため、データの真正性を確保することは極めて重要です。データが真正でなかったり操作されていたりすると、モデルが誤ったパターンを学習し、パフォーマンスの低下や予測に基づく有害な意思決定につながる可能性があります。データの真正性は、教育、ヘルスケア、金融サービス、小売、製造、政府サービス、サイバーセキュリティなど、AIモデルに基づく意思決定がさまざまな業界に重大な影響を及ぼす分野で特に重要です。

さらに、AIアプリケーションのために個人データを収集・処理する際には、適切なデータ同意を取得し、一般データ保護規則（GDPR）などの規制を遵守することが不可欠です。GDPRは、組織が個人データを収集・処理する前に、個人から明示的な同意を得ることを義務付けており、また、個人に対してデータへのアクセス、修正、消去の権利を認めています。

- **評価基準**：真正性を監査されたデータの割合を定期的に測定し、一定期間で100%の検証を目指します。さらに、データ同意とGDPR規制の遵守状況を監視します。

- 可能な限り、個人は自分に関するデータを訂正する権利を持つべきです。例えば保険ではFTCがこれを求めています：FTCはAI規制を追求し、偏ったアルゴリズムを禁止します。
- **RACIモデル**：データ管理チーム（Responsible）、最高データ責任者（Accountable）、法務担当者およびコンプライアンス部門（Consulted）、セキュリティチーム（Informed）。
- **ハイレベルの実施戦略**：データの真正性を定期的に監査するためのポリシーを導入します。さらに、データ同意の取得、データプライバシーの確保、GDPR規制の遵守のためのプロセスを確立します。
- **継続的なモニタリングと報告**：検証されたデータの割合に関する定期的な報告や、真正性、不正なデータ変更、データの取得や利用についての同意およびGDPR規制の遵守。

データの真正性を確保し、適切なデータ利用への同意を取得し、GDPRのような規制を遵守することで、組織は個人データに対する個人のプライバシーと権利を尊重しながら、信頼できるAIモデルを構築することができます。

12 匿名化と仮名化

匿名化と仮名化は、信頼できるAIシステムにおける個人データのプライバシーを以下のように保護します：

- 匿名化によって、データから識別子が永久に取り除かれます。
  個人を再特定することをできなくし、データ保護法の遵守を支援します。GDPRなど多くの場合、匿名化されたデータはデータ保護規制の範囲から除外されます。
- 仮名化は、識別子をシステムが生成した識別子または人工的な識別子に置き換えます。
  個人は自分のデータとリンクしたままだが、本当の身元は保護されます。仮名化は、GDPRやHIPAAのような特定のデータ保護規制の下で要求されています。
- 評価基準：匿名化および仮名化技術により、特定可能な個人データを99%削減することを目標とします。直接識別子（フルネーム、SSN＝ソーシャルセキュリティナンバー、クレジットカード番号など）と準識別子を区別することが重要です。直接識別子は、高い確実性で個人を特定するために使用できるため、より厳格な削減措置が必要となります。さらに、クレジットカード番号のような直接識別子は盗難につながる可能性があり、SSNのようなものはなりすましにつながる可能性があります。準識別子（年齢、郵便番号、性別）は、プライバシーにとって重要ではありますが、データ保護とユーザビリティのバランスを確保するため、それほど厳密な削減は行われないかもしれません。とはいえ、多くの準識別子はAIにバイアスをかける可能性があります（年齢、場所、性別、人種、セクシュアリティなど）。さらに、特定の準識別子は、宗教的信条、政治的所属、性的指向、民族的出身など、特別な注意を要する個人データの特別なカテゴリー（GDPR第9条-個人データの特別なカテゴリーの処理）に該当する可能性があります。

  準識別子はまた、個人を再識別するために組み合わされる可能性もあります。準識別子は、特にヘルスケアやマーケティングのような分野では、AIシステムの使いやすさや機能性にとって極めて重要である一方、再識別のリスクもあります。異なるデータセットが組み合わされた場合、たとえ匿名化または仮名化されていたとしても、これらの準識別子は個人を再識別するために調整される可能性があります。

例えば、匿名化された医療記録を含むデータセットと、有権者登録記録などの一般に利用可能なデータセットを組み合わせることができます。両方のデータセットに詳細な人口統計学的情報が含まれていれば、これらの準識別子に基づいて記録を照合することが可能になり、匿名化されたデータセットの個人を再識別することができる可能性があります。

このリスクを軽減するためには、バランスの取れたアプローチが必要です。特にデータ処理技術が進化し、より洗練されるにつれて、再識別化のリスクを継続的に評価することが不可欠です。再識別を防ぐために統計的ノイズをデータに加える差分プライバシーのような高度な技術を採用することで、プライバシー保護をさらに強化することができます。
さらに、データの匿名化および仮名化のプロセスが、進化するデータ保護法および規制に適合していることを確認するためには、定期的な監査とコンプライアンスチェックが不可欠です。

- **RACI**モデル：データ管理者(Responsible)、データ保護責任者(Consulted)、最高プライバシー責任者
  オフィサー（Accountable）、法務チーム（Consulted）、ITチーム（Informed）、セキュリティチーム（Consulted）、データガバナンスチーム（Consulted）、データサイエンティスト（Informed）。
- ハイレベルな実装戦略： 最先端の匿名化と仮名化技術には、差分プライバシー、同型暗号化、セキュアなマルチパーティ計算などの高度な暗号化手法を活用し、機密データを保護しながら分析の有用性を維持することが含まれます。例えば、企業はk-匿名性、l-多様性、t-Closenessのような技術を採用することで、データセット内で個人の身元を確実に隠しながら、意味のある分析を行うことができます。さらに、トークン化やデータマスキングなどの技術を採用することで、センシティブなデータを非センシティブな同等データに置き換え、プライバシー保護をさらに強化することができます。
- 継続的なモニタリングと報告：これらの技術の有効性を定期的に評価します。
- アクセス制御のマッピング：冗長化または匿名化ルールおよび非仮名化ツールへのアクセスを制限します。
- 基礎的なガードレール：データの匿名化および仮名化に関する一般データ保護規則（GDPR）のガイドラインおよびクラウドベースのサービスにおけるトークン化のガイダンスに従います。

1.3 データ最小化

データ最小化とは、特定の目的または機能を達成するために必要な量のデータのみを使用することを指します。この実践は、GDPRなど多くのデータ保護規制における要件です。また、匿名化されたデータの再識別化を防止するために利用できる手法の一つでもあります。この手法により、収集、保存、ML目的のために使用されるデータの量と種類が制限されます。この実践は、データ主体のプライバシーとセキュリティの保護に役立ち、MLモデルの性能と効率を向上させます。MLでは、データの最小化は、モデルの学習と性能に不可欠な特徴やデータポイントを注意深く選択する一方で、無関係なデータや過剰なデータを除外することを含みます。これは、信頼できるAIの基礎となる説明可能性、公平性、透明性、プライバシーの柱と関連しています。

- **評価基準**：組織の事業目的と責任に基づき、収集される非本質的なデータを少なくとも適切な割合で減少させることを目指します。

- **RACIモデル**：データ収集チーム（Responsible）、データプライバシー室（Consulted）、データガバナンス委員会（Accountable）、コンプライアンスチーム（Consulted）、セキュリティチーム（Informed）、データサイエンティスト（Informed）。
- **ハイレベルな実装戦略**：最小限のデータ取得に焦点を当てた厳格なデータ収集ガイドラインを作成。
- **継続的なモニタリングと報告**：収集したデータ量を追跡し、その必要性を評価。
- **アクセス制御のマッピング**：誰が追加のデータ収集を許可できるかを制御。
- **基礎的なガードレール**：GDPRのプライバシー・バイ・デザインの原則を採用。

1.4 データへのアクセス制御

機械学習におけるデータへのアクセス制御は、MLモデルのトレーニング、テスト、デプロイに使用されるデータにアクセスし、相互作用できる者を管理し制限することを含みます。このプロセスは、許可された個人またはシステムのみがデータを閲覧、修正、使用できることを保証します。効果的なアクセス制御は、機密情報を保護し、データの整合性を維持し、プライバシー規制を遵守するために、ML環境において極めて重要です。一般的には、ユーザーの身元を確認する認証メカニズム、ユーザーの役割に基づいて特定のアクセス権を付与する認可プロトコル、データへのアクセスと使用状況を追跡する監査システムが含まれます。

一般的に組織では、AIモデルはさまざまなデータソースやシステムから集約されたデータで実行されます。そのためAIモデルは、基礎となるデータソースシステムのアクセス制御定義とポリシーを尊重する必要があります。つまり、AIモデルは、データ管理者やシステム管理者によって定義されたアクセス制御ルールと権限に基づいて、処理することが許可された特定のデータにのみアクセスできるようにする必要があります。適切なアクセス制御を維持することで、AIインフラストラクチャ全体のデータプライバシー、セキュリティ、および規制要件へのコンプライアンスが確保され、AIモデルによる機密データや機密性の高いデータへの不正アクセスや誤用が防止されます。

- **評価基準**：年間0.5%未満の不正データアクセスインシデントの達成
- **RACIモデル**：セキュリティチーム（Responsible）、最高情報セキュリティ責任者（Accountable）、データガバナンス機関、データ管理者、IT チーム（Consulted）、運用チーム（Informed）。
- **ハイレベルの実装戦略**：アクセス制御のハイレベル実装戦略の一環として、レイヤード・セキュリティ・モデルを実装します。このモデルは、堅牢な認証・認可プロトコルだけでなく、多要素認証（MFA）、役割ベースのアクセス制御（RBAC）、最小特権の原則（PoLP）などの先進技術も統合する必要があります。
- **継続的な監視と報告**：アクセスログを監視し、監査を実施します。重要なモデルや生成されたモデル、データへのリスクベースのアクセスにフラグを立てることができるツールを使用します。
- **アクセス制御のマッピング**：アクセス許可を定期的に監視・管理します。
- **基礎的なガードレール**：ISO/IEC 42001、ISO/IEC 27001、ISO/IEC 27701、NIST 800-53、およびOWASP Top 10 A07:2021-Identification and Authentication Failures に従ったベストプラクティスを実施します。

1.5 安全な保管と送信

機械学習では、機密データを保護するために、安全な保管と送信が重要です。セキュアな保管には、不正アクセスを防ぐために保存状態のデータを暗号化し、強固なアクセス制御を採用し、定期的なセキュリティ監査を実施することが含まれます。セキュアな送信では、トランスポートレイヤーセキュリティ（TLS）やフィールドレベル暗号化、エンベロープ暗号化などのプロトコルを使用して、送信中のデータを暗号化します。これにより、システム間やネットワーク間でデータが転送されている間も、機密性が保たれ、無傷であることが保証されます。

これらの慣行は、データの不正アクセス、不正使用、不正開示、悪意または偶発的なデータの変更、削除、破損を防止することにより、データとMLモデルのセキュリティを強化します。

- 評価基準：転送中および保存中のすべてのデータに対し、256ビットAES以上の暗号化基準を維持します。
- **RACI**モデル：セキュリティチーム（Responsible）、最高情報セキュリティ責任者（Accountable）、コンプライアンス及び法務チーム（Consulted）、経営陣（Informed）。
- ハイレベルな実装戦略：高度な暗号化技術と、政策によるAIデータとモデルの自動削除に投資します。
- 継続的なモニタリングとレポーティング：リアルタイムのセキュリティ監視のためのツールを使用します。
- アクセス制御のマッピング：安全な保存と転送をアクセス制御と統合します。
- 基盤となるガードレール：NIST（National Institute of Standards and Technology：米国国立標準技術研究所）のガイドラインとプライバシー法に従い、転送中および保存中のデータを保護します。

2. モデルセキュリティ

モデルセキュリティは、幅広いコンポーネントを包含する多面的なタスクです。これらには、モデルAPIのアクセス制御、認証と認可のフレームワーク、レート制限、モデルライフサイクル管理、セキュアなモデルランタイム環境、ハードウェアベースのセキュリティ機能、ネットワークセキュリティ制御、OSのセキュリティ強化、セキュアな設定、コンテナとクラウド環境におけるセキュリティなどが含まれます。これらの重要な領域それぞれについて、評価基準を検討し、RACI モデルを使用して責任を割り当て、ハイレベルの実装戦略の概要を示し、継続的な監視と報告のメカニズムを確立し、アクセス制御をマッピングし、NIST AI RMF、NIST SSDF、NIST 800-53、CSA CCM などの標準の基礎的なガードレールを参照します。

2.1. モデルへのアクセス制御

アクセス制御は、AIモデルの安全性を確保し、許可された人員とシステムのみが機密データや機能とやり取りできるようにする上で、極めて重要です。AIモデルガバナンスの領域では、アクセス制御手段は堅牢で適応性があり、組織や業界のセキュリティ標準に沿ったものでなければなりません。
認証や認可のフレームワークからレート制限やライフサイクル管理まで、AIモデルの完全性は、強力で柔軟なアクセス制御プロトコルにかかっています。これらのプロトコルは、誰が、いつ、どのような状況でAIモデルにアクセスできるかを規定します。組織がAI導入の複雑さを乗り越えるにつれ、リスクを軽減し、知的財産を保護し、規制上のコンプライアンス基準を守るために、包括的なアクセス制御戦略の導入が不可欠になっています。さらに、AIモデルのアクセス制御を組織の既存のセキュリティフレームワークと統合し、システム全体のレジリエンスを高めることも重要です。これには、誰がどのような状況でAIモデルと対話できるかを規定するきめ細かなアクセスポリシーを確立することが含まれます。さらに、多要素認証やロールベースアクセス制御など、堅牢な認証メカニズムを導入することで、AI システムのセキュリティ状態をさらに強化することができます。定期的な監査とアクセスログの監視は、不正アクセスの試みを迅速に検知し対応するために不可欠です。AIモデルのアクセス制御を既存のセキュリティフレームワークと緊密に統合することで、組織はサイバー脅威に対する防御を強化し、AIシステムとデータの完全性と機密性を確保することができます。

2.1.1 認証と認可のフレームワーク

機械学習モデルのための認証と認可のフレームワークは、MLモデルと関連データへのアクセスが厳密に制御・管理されることを保証する、セキュリティにとって不可欠なものです。認証は、多くの場合、パスワード、トークン、生体認証などの方法を使用して、ユーザーまたはシステムの身元を確認します。同時に認可は、アクセスレベルを決定し、確立された役割と権限に基づいて、誰がモデルを閲覧、編集、または使用できるかを定義します。これらのフレームワークは、機密情報を保護し、データの完全性を保持し、プライバシーとセキュリティの規制を遵守するために不可欠であり、それによってMLモデルとそのデータへの不正アクセスや変更を防止します。AIにおいて特に重要な検証は、あらゆるユーザーとエンティティが目的とコンテキストに基づいて、AIデータとモデルの適切な使用が承認されることです。この側面はAIのデジタル権利に組み込まれ、アクセスと認可の基礎となります。

- **評価基準**： API 経由でアクセスされない AI モデル（1.2.1 で説明）において、認証と認可のフレームワークを 100% カバーする。
- **RACIモデル**：セキュリティチーム（Responsible）、最高情報セキュリティ責任者（Accountable）、法務チーム（Consulted）、AI開発チーム（Informed）。
- **ハイレベルな実装戦略**：AIモデルへの安全なアクセスのための包括的なフレームワークを開発し、実施する。
- **継続的な監視と報告**：認証および認可メカニズムを定期的に監査する。
- **アクセス制御のマッピング**：モデル固有の要件に基づいてアクセスをカスタマイズする。
- **基礎的なガードレール**：リスク管理には、NIST 800-207、NIST 800-53、NIST SP 800-63、NIST AI RMF を使用する。

21.2. モデルインターフェースレート制限

機械学習(ML)におけるモデルインターフェースのレート制限は、ユーザーやシステムが与えられた時間内にMLモデルに対して行えるリクエストの数を制限することを必要とします。この方法は、モデルの負荷を管理し、（DoS攻撃などの）侵害を防ぎ、ユーザー間の公平なリソース配分を保証するために極めて重要です。レート制限は、APIやウェブインターフェースなど、ユーザがMLモデルとやりとりする様々なインターフェースレベルで実装することができます。リクエストレートを制御することで、モデルのパフォーマンス、安定性、可用性を維持することができ、高い負荷や潜在的な攻撃シナリオの下でも、効率的かつ確実に動作し続けることが保証されます。

- **評価基準**：サービス拒否（DoS）または分散型サービス拒否（DDoS）攻撃によるダウンタイムを削減する。
- **RACIモデル**：プラットフォームサポートチーム（Responsible）、ソリューションオーナー（(Accountable）、データサイエンティスト（Consulted）、リスク管理チーム（Consulted）、AIモデルユーザー（Informed）。
- **ハイレベルな実装戦略**：AIモデルインターフェースの使い過ぎや乱用を防ぐため、レート制限を実施する。
- **継続的なモニタリングとレポート**：利用パターンを追跡し、それに応じてレート制限を調整する。
- **アクセス制御のマッピング**：ユーザーベースのレート制限戦略を実装する。
- **基礎的なガードレール**：OWASP LLM 04: モデルサービス拒否に従う。

21.3. モデルライフサイクル管理におけるアクセス制御

機械学習（ML）のモデルライフサイクル管理におけるアクセス制御は、MLモデルのライフサイクル（開発、デプロイ、メンテナンスの各段階）を通して、MLモデルへのアクセスと相互作用を管理・規制することを含みます。このプロセスにより、さまざまな段階で、許可された担当者やシステムのみがMLモデルと対話できることが保証され、それによってモデルの完全性やパフォーマンスの問題を引き起こす可能性のある不正アクセスや変更からモデルが保護されます。堅牢なアクセス制御の実装は、MLモデルのセキュリティと有効性を維持するために極めて重要です。これは、潜在的なデータ漏洩やモデルの悪用を防止し、データプライバシーとセキュリティの規制への準拠を保証するのに役立つからです。モデルのライフサイクルの各段階で慎重にアクセス制御を行うことで、組織は安全で効率的なML開発環境を育成しながら、ML資産を保護することができます。モデルライフサイクル管理におけるアクセス制御は、データへのアクセスや使用に関する一貫した明確なポリシーと手順を提供し、データの出所と履歴を文書化して

記録することにより、MLモデルとデータの透明性とアカウンタビリティを向上させます。

この領域は、信頼できるAIの基本的な柱であるプライバシー、透明性、説明責任につながります。

- **評価基準**：AIモデルとデータへのアクセスが、モデルライフサイクルの全段階において、許可されたユーザーとシステムに制限されていることを確認する。
- **RACIモデル**：AIモデル ガバナンス・チーム（Responsible）、最高データ責任者（Accountable）、セキュリティチーム、法務チーム、コンプライアンスチーム（Consulted）、運用スタッフ（Informed）
- **ハイレベルな実装戦略**：
    - データとモデルを機密レベルに基づいて分類する。
    - モデルのライフサイクルの各フェーズごとに、ユーザーロールにマッピングされた個別のアクセス制御ルールを定義する。
    - 既存のIAM（アイデンティティおよびアクセス管理）ソリューションとアクセス制御を統合する。
    - すべてのアクセス要求とデータ/モデルの使用状況を記録し、監査する。
- **継続的なモニタリングと報告**：
    - 不正アクセス試行に対するアラートを送信する。
    - ユーザーアクセスのレビューと再認証を行う。
    - アクセスと管理の監査を定期的に実施する。
    - 不審なアクティビティに対するアラートしきい値を設定する。
- **アクセス制御のマッピング**：
    - 開発段階：データサイエンティスト、MLエンジニアへのアクセスを制限する
    - テスト段階：品質保証チームのアクセスを追加する
    - 生産段階：本番システムへの厳重に管理されたアクセスを許可する
- **基礎的なガードレール**
    - NIST 800-53、NIST AI RMF フレームワークなどの標準との整合性
    - CSA CCM のようなベストプラクティス・フレームワークに照らして、管理策を検証する。

22. セキュアモデルランタイム環境

セキュアなAIモデルのランタイム環境のためにレジリエントシステムを構築するには、堅牢なハードウェア、ネットワーク、およびソフトウェアのセキュリティ管理を統合する必要があります。進化する脅威からAIのデプロイメントを保護することに重点を置き、企業は、完全性・機密性・可用性を維持するために、ランタイム環境を綿密に設計し、強化します。信頼された実行環境を活用するハードウェアベースのセキュリティ機能から、ファイアウォールやセグメンテーションのようなネットワークセキュリティ制御まで、各コンポーネントは、深レイヤー防御戦略の基本構造に複雑に織り込まれています。NIST AI RMF、NIST 800-53、CSA CCMなどの業界標準に準拠するための確固たるコミットメントにより、チームはこれらの重要なセキュリティ対策の実装、監視、管理を指揮すべく、分野を超えて協力します。

221. ハードウェアベースのセキュリティ機能

機械学習（ML）モデル向けのハードウェアベースのセキュリティ機能には、MLアプリケーションのセキュリティを強化するコンピューティングハードウェアの物理的およびアーキテクチャ的要素が含まれます。これには、信頼された実行環境（Trusted Execution Environments：TEE）、分離された安全な処理のためのコンフィデンシャルコンピューティング、機密コードとデータを保護するセキュアエンクレーブ、安全な暗号化操作のためのハードウェアセキュリティモジュール（Hardware Security Modules：HSM）、信頼されたソフトウェア初期化を保証するセキュアブートメカニズム、不正な物理アクセスを防止する物理的耐タンパーメカニズム が含まれます。これらの機能は、セキュリティの基礎となるレイヤーを提供する上で不可欠であり、MLモデルが機密データを扱い、改ざんや不正アクセスなどのさまざまな脅威から強固に保護する必要がある金融、医療、防衛などの重要な分野では特に重要です。

- **評価基準**：企業で定義した割合の該当する AI システムの、ハードウェアベースのセキュリティ機能を使用している。
- **RACIモデル**：ハードウェアセキュリティチーム（Responsible）、最高技術責任者（Accountable）、調達部門（Consulted）、システム管理者（Informed）。
- **ハイレベルの実装戦略**：信頼された実行環境（NVIDIAのコンフィデンシャル・コンピューティング・アプローチなど）、GPU、TPU、その他のハードウェア・セキュリティ対策をAIシステムに統合する。
- **継続的なモニタリングとレポート**：ハードウェアのセキュリティの完全性と機能性を定期的にチェックする。
- **アクセス制御のマッピング**：許可された担当者のみがハードウェアのセキュリティ設定にアクセスできるようにする。
- **基礎的なガードレール**：NIST AI RMF および NIST 800-53 ガイドラインを実施する。

222　ネットワークセキュリティコントロール

機械学習（ML）モデルのためのネットワークセキュリティコントロールは、MLモデルとその関連データをネットワークベースの脅威や脆弱性から保護するために実装された対策とプロトコルです。ゼロトラストアーキテクチャを採用し、AIシステムをより広範なネットワークから分離します。AIシステムとMLモデルを隔離することで攻撃サーフェスを減らし、攻撃者がネットワーク内でラテラルムーブメントを困難にします。これらの制御は、MLモデルで使用される送信中のデータを保護し、不正アクセスを防止し、ML通信の完全性と機密性を確保するために不可欠です。主なネットワークセキュリティ管理には、送受信ネットワークトラフィックを監視・制御するための次世代ファイアウォール、転送中のデータを保護するためのTLSなどの暗号化プロトコル、侵入検知・防御システム（IDPS）、攻撃を特定・軽減するためのウェブアプリケーションファイアウォール（WAF）、安全な通信チャネルを構築するための仮想プライベートネットワーク（VPN）、MLモデルへのAPIコールを管理・認証するための安全なAPIゲートウェイの使用が含まれます。これらの対策は、データとモデルのセキュリティが最も重要なML環境、特にモデルがインターネットやクラウド環境を含むネットワーク経由でアクセスまたは管理される場合に極めて重要です。

- **評価基準**：AIシステム全体のネットワークセキュリティポリシーに100%準拠する。
- **RACI モデル**：ネットワークセキュリティチーム（Responsible）、最高情報セキュリティ責任者（Accountable）、IT 運用部門（Consulted）、全ネットワークユーザ（Informed）。
- **ハイレベルな実装戦略**：ファイアウォールやセグメンテーションなど、包括的なネットワークセキュリティ対策を実施する。
- **継続的なモニタリングとレポート**：ネットワークの出入口トラフィックとインフラ内のトラフィックを定期的に監視し、制御のコンプライアンスを強化する。定期的な侵入テストと脆弱性評価を取り入れることは、ネットワークセキュリティ管理の潜在的な弱点が悪用される前に特定するための積極的なアプローチであり、ML モデルとそのデータの強固な保護を保証する。
- **アクセス制御のマッピング**：ネットワークアクセス制御を特定の役割やモデル要件に合わせて調整する。
- **基礎的なガードレール**：CIS Controls V8 related to Network Security

223. OSレベルのセキュリティ強化とセキュアなコンフィギュレーション

機械学習（ML）モデルのためのOSレベルのセキュリティ強化とセキュアな設定には、リスクを軽減して脆弱性を減らすために、MLモデルとアプリケーションが実行される基盤となるオペレーティングシステム（OS）を強化することが含まれます。OSはこれらのアプリケーションの基盤レイヤーを形成するため、このプロセスはML運用のためのセキュアな環境を構築する上で極めて重要であ。重要な点は以下の通りです：

定期的なアップデートとパッチ管理：既知の脆弱性から保護するために、OSとそのコンポーネントを最新のセキュリティパッチとアップデートで最新の状態に保つ。

最小限のインストール：MLの運用に必要のないOSの不要なサービス、アプリケーション、機能を削除または無効にし、潜在的な攻撃対象領域を最小限に抑える。

セキュリティ設定の構成：ファイアウォールの有効化、ユーザー権限の設定、システムへのアクセスや使用の方法を規定するセキュリティポリシーの実装など、セキュリティを強化するためにOSの設定を調整する。

ユーザーアクセス制御：厳格なユーザーアクセス制御を実施し、許可されたユーザーのみがMLシステムにアクセスできるようにし、最小権限の原則を適用する。ここでは、ユーザーは各自のタスクを実行するために必要なアクセス権のみが付与される。

監視と監査：監視・監査ツールを設定し、OSのアクティビティや変更を追跡することで、セキュリティインシデントの検出と対応に役立てる。

安全な通信プロトコル：MLシステムとのすべての通信が暗号化され、安全であることを保証する。

これらの対策は、MLシステムとそのデータの完全性・機密性・可用性を損なう可能性のある様々な脅威からMLシステムを保護し、MLシステムの強固なセキュリティポスチャを構築するのに役立ちます。

- **評価基準**：AIシステムの100%が、堅牢化されたOSとセキュアな設定で稼働することを保証する。
- **RACIモデル**：システム管理チーム（Responsible）、最高情報セキュリティ責任者（Accountable）、セキュリティチーム（Consulted）、エンドユーザー（Informed）。
- **ハイレベルな実装戦略**：OSのセキュリティ強化とセキュアな構成設定のベストプラクティスを適用する。
- **継続的なモニタリングとレポート**：コンプライアンスとセキュリティ脅威に対する脆弱性を監視する。
- **アクセス制御のマッピング**：システム設定を変更できるユーザーを制限する。
- **基礎的なガードレール**：NIST 800-53、CIS、DISA STIGs Benchmarksを安全な設定に活用する。

224. K8sとコンテナのセキュリティ

機械学習（ML）のためのKubernetes（K8s）とコンテナのセキュリティは、コンテナ化されたMLアプリケーションとそのデプロイ環境のセキュリティを確保するために設計されたプラクティスとツールのセットを指します。コンテナオーケストレーションプラットフォームであるKubernetesとコンテナ化技術は、MLモデルとワークロードのデプロイと管理に広く使用されています。この文脈でのセキュリティには、コンテナが安全に構成され、優れたリスク管理フレームワークの下で脆弱性が低減されていること、堅牢なKubernetesクラスタセキュリティ（ネットワークポリシー、アクセス制御、ポッドセキュリティを含む）を実装すること、Kubernetes環境内の通信チャネルの保護を確認することが含まれます。また、コンテナ特権の管理、脆弱性の定期的なスキャン、MLワークフローにおけるコンテナの実行方法と相互作用を管理するポリシーの実施、コンテナ化されたデプロイ環境における不正アクセス・侵害・その他のセキュリティ脅威からのMLモデルとデータの保護も含まれます。

- **評価基準**：AIシステム内のコンテナ環境におけるセキュリティ侵害率の低減を目標とする。
- **RACIモデル**：コンテナ管理チーム（Responsible）、CTO（Accountable）、セキュリティチーム（Consulted）、DevOpsチーム（Consulted）、アプリケーション開発チーム（Informed）。
- **ハイレベルの実装戦略**：コンテナ環境におけるAIアプリケーションの安全なデプロイメントを保証する。
- **継続的なモニタリングとレポーティング**：コンテナオーケストレーションツールのセキュリティ評価を定期的に実施する。
- **アクセス制御のマッピング**：コンテナ環境に対する厳格なアクセスポリシーを定義し、実施する。
- **基礎的なガードレール**：OWASP Kubernetes Top Ten、NIST SSDF、CNCF Security Whitepaper、CIS Benchmarks、NIST 800-190、NIST 800-53のベストプラクティスに従う。

225. クラウド環境セキュリティ

機械学習（ML）モデルのためのクラウド環境セキュリティは、クラウドベースのインフラにおけるMLモデルとその関連データを保護するために実装される戦略と対策を含みます。これには、クラウド内でのデータの保存と処理の保護、クラウドリソースへのアクセス制御の管理、静止時および転送時のデータの暗号化、クラウドプラットフォーム上にデプロイされたMLモデルのセキュリティの確保などが含まれます。また、定期的な脆弱性評価、クラウド固有のセキュリティ基準への準拠、アイデンティティとアクセス管理のベストプラクティスの実施も含まれます。このセキュリティは、不正アクセス・データ漏洩・その他のサイバー脅威を防止し、クラウド環境の動的かつ分散的な性質におけるMLモデルとデータの完全性と機密性を確保するために極めて重要です。

- **評価基準**：AIデプロイメントにおけるクラウドセキュリティポリシーの100%遵守に努める。
- **RACIモデル**：クラウドセキュリティチーム（Responsible）、最高情報セキュリティ責任者（Accountable）、ITガバナンス（Consulted）、クラウドサービス利用者（Informed）。
- **ハイレベルの実装戦略**：AIシステムに強固なクラウドセキュリティ対策を導入する。
- **継続的なモニタリングとレポーティング**：クラウドに特化した監視ツールを使用して、脅威を検出し、アラートを出す。
- **アクセス制御のマッピング**：クラウドベースのAIアプリケーションのアクセス制御をカスタマイズする。

- **基礎的なガードレール**：CSA CCM を採用する。クラウドネイティブなアプリケーション保護プラットフォーム（CNAPP）を採用し、優先する。

2.3 脆弱性とパッチ管理

以下に示す責任とアプローチは、組織の広範なセキュリティと開発のプロセスに統合されるべきであり、AIシステムが安全かつ効果的に開発、デプロイ、維持されることを保証します。これらのプロセスの定期的な見直しと更新は、AIシステムを攻撃、エラー、または障害にさらす可能性のあるソフトウェアの欠陥や弱点を特定し、優先順位を付け、修正を適用する際に、進化する脅威や技術に適応するのに役立ちます。以下に示す技術は、AIシステムのセキュリティ、確実性、および信頼性を確保するのに役立ちますので、AI組織の責任の一側面となります。

2.3.1 MLコードの完全性保護

MLコードの完全性保護は、機械学習（ML）アプリケーションで使用されるソースコードのセキュリティと完全性を確保するために採用される対策と実践を指しています。これには、不正な変更からコードを保護し、その真正性を確保し、開発とデプロイのプロセスを通してその品質を維持することが含まれます。主なプラクティスには、変更を追跡・管理するためのバージョン管理システムの導入、コードの信頼性を検証するためのコード署名の使用、脆弱性を検出するための定期的なコードレビューと監査の実施、潜在的なセキュリティ問題を特定するための静的・動的コード解析ツールの採用などがあります。これらの保護は、MLアプリケーションの信頼性を維持し、悪意のあるコードインジェクションを防止し、MLモデルが侵害されることなく意図されたとおりに実行されることを保証するために不可欠です。

- **評価基準**
  - 完全性保護（integrity protection）が適用されるMLコードの割合
  - 完全性チェックの頻度
  - 重要な脆弱性に対処した割合
- **RACI Model**：AI開発チーム（ Responsible）、AI開発マネージャー（Accountable）、セキュリティチーム、DevOpsチーム（Consulted）、コンプライアンスチーム（Informed）。
- **ハイレベルの実装戦略**：
  - MLモデルに対する自動化された完全性チェックの実装。
  - CI/CDパイプラインへの完全性チェックの組み込み。
  - MLコードに対する定期的な脆弱性評価の確立。
  - MLアルゴリズムに特化したセキュアなコーディングプラクティスの実施。
  - MLモデルの異常動作のランタイムモニタリングの実装。
- **継続的なモニタリングと報告**：
  - モニタリングツールを活用して、MLモデルの動作の異常を検出。
  - 検出された異常に対するインシデント対応手順を確立。
  - 完全性チェックの結果や発見された異常を定期的に報告。
- **アクセス制御のマッピング**：
  - 役割と責任に基づいてMLコードリポジトリへのアクセスを許可。
  - MLコードへのアクセスに最小特権原則を導入。

    - 機微なMLコードのリポジトリに多要素認証を利用。

- 基礎的なガードレール
    - MLシステムをセキュアにするガイドラインについてNIST AI RMFを参照。
    - NIST800-53で推奨されているセキュリティ管理策のうち、MLシステムに関連するものを実施。
    - クラウド固有のセキュリティに関する考慮事項については、CSA CCM に従う。

232 MLトレーニングとデプロイのコードのためのバージョン管理システム

MLトレーニングおよびデプロイのコードのバージョン管理システムは、機械学習（ML）プロジェクトを管理する上で不可欠なツールであり、チームがMLモデルのトレーニングおよびデプロイに使用するコードベースの変更を追跡および管理できるようにします。これらのシステムは、変更の履歴を管理し、変更を簡単に追跡できるようにし、必要に応じて以前のバージョンへのロールバックをサポートすることで、開発者やデータサイエンティスト間のコラボレーションを促進します。バージョン管理システムは、様々なバージョンのMLモデルと関連するデータセットを扱う上で非常に重要であり、ML試験とデプロイにおける一貫性と再現性を保証します。バージョン管理を使用することで、チームは開発からテスト、デプロイ、メンテナンスに至るまで、MLモデルのライフサイクルを効率的に管理することができます。

- **評価基準**
  - バージョン管理下のコードの割合
  - コミットとアップデートの頻度
  - バージョン管理ポリシーのコンプライアンス
- **Responsibility (RACI Model)** :
  - Responsible: 開発チーム
  - Accountable: 開発マネージャー
  - Consulted: DevOpsチーム、QAチーム
  - Informed: セキュリティチーム
- **ハイレベルの実装戦略** :
  - 集中型バージョン管理システム（Gitなど）の導入
  - ブランチとマージのポリシーの実施
  - マージ前のコードレビューとチェックの自動化
  - リリース管理のためのバージョンタグの導入
- **継続的なモニタリングと報告** :
  - コミット活動を監視し、不規則なパターンを特定
  - 承認されていない変更や異常な活動に対するアラートの設定
  - コンプライアンスのためのバージョン管理ログの定期的なレビュー
- **アクセス制御のマッピング** :
  - ロールに基づくリポジトリのためのアクセス制御リストの定義
  - リポジトリへのアクセスに二要素認証を導入
  - リポジトリへのアクセスを定期的に監査し、コンプライアンスを確保
- **基礎的なガードレール**
  - セキュアなソフトウェア開発ガイドラインのためにNIST SSDFに従うこと。
  - バージョン管理および変更管理に関連する、NIST 800-53 に概説されている管理策を実施。

2.3.3 承認されたバージョンを検証するためのコード署名

機械学習（ML）の文脈における、承認されたバージョンを検証するためのコード署名は、MLモデルで使用されるソフトウェアコードの真正性と完全性を検証するためにデジタル署名が使用されるセキュリティ慣行です。このプロセスでは、通常、デプロイのためのレビューと承認が行われた後に、暗号署名をコードに添付します。署名は、コードが署名された後に変更されたり改ざんされたりしていないことを検証するシールの役割を果たします。MLワークフローにおいて、コード署名は、MLモデルのトレーニング、テスト、デプロイに使用されるコードが、正確で承認されたバージョンであり、悪意を持って変更されていないことを保証するために重要です。このプラクティスは、特にモデルが異なる環境に分散されたり、複数のチームや組織で共有されたりする場合に、MLソフトウェアのサプライチェーンの信頼性を維持するのに役立ちます。

- **評価基準**
  - 承認された証明書で署名されたコードの割合
  - コード署名ポリシーのコンプライアンス
  - コード署名チェックの頻度
- **RACI Model** :
  - Responsible: 開発チーム
  - Accountable: 開発マネージャー
  - Consulted: セキュリティチーム、リリース管理チーム
  - Informed: コンプライアンスチーム
- **ハイレベルの実装戦略** :
  - ビルドプロセス中にコード署名を実装
  - コード署名証明書を安全に管理
  - デプロイ前のコード署名チェックを自動化
  - リプレイ攻撃を防ぐために、署名されたコードにタイムスタンプを実装
- **継続的なモニタリングと報告** :
  - コード署名のアクティビティと証明書の使用状況を監視
  - 不正なコード署名の試みに対するアラートの実装
  - コンプライアンスのためにコード署名ログを定期的にレビュー
- **アクセス制御のマッピング** :
  - コード署名インフラへのアクセスを許可された担当者のみに制限
  - コード署名証明書の役割ベースアクセス管理(RBAC)の実装
  - コード署名インフラのアクセスログを定期的に確認
- **基礎的なガードレール**
  - コード署名と完全性チェックに関連する NIST 800-53 管理を参照すること。

2.3.4 Infrastructure as Codeアプローチ

機械学習(ML)モデルのためのInfrastructure as Code(IaC)アプローチでは、物理的なハードウェア構成や対話的な構成ツールではなく、機械が読み取り可能な定義ファイルを通じてコンピューティングインフラを管理し、プロビジョニングします。この手法により、サーバー、ストレージ、ネットワーキングリソースなど、MLモデルに必要なインフラのセットアップ、構成、管理を一貫性のある反復可能な方法で自動化することができます。IaCにより、MLチームは、最小限の手動介入で、多様な環境（クラウド、オンプレ

ミス、ハイブリッドセットアップなど）にモデルを迅速に展開し、拡張することができます。このアプローチは、MLモデルのインフラ展開の効率性と信頼性を高め、インフラ状態が保守可能で、バージョン管理され、定義された標準に準拠していることを保証し、MLプロジェクトにおけるコラボレーションの改善と運用リスクの低減につながります。

- **評価基準**
  - コードとして管理されるインフラの割合
  - IaCベストプラクティスのコンプライアンス
  - インフラの更新と見直しの頻度
- **RACIモデル**:
  - Responsible: DevOpsチーム
  - Accountable: DevOpsマネージャー
  - Consulted: 開発チーム、セキュリティチーム
  - Informed: オペレーションチーム
- **ハイレベルな実装戦略**:
  - TerraformやAWS CloudFormationなどのIaCツールの活用
  - インフラコードのバージョン管理
  - インフラ変更のテストと検証の自動化
  - インフラのドリフトの検出と修復
- **継続的なモニタリングと報告**:
  - インフラの変更とドリフトを監視
  - 未承認または予期せぬ変更に対するアラートの実装
  - インフラコードが標準に準拠しているかの定期的な見直し
- **アクセス制御のマッピング**:
  - 役割に基づいてインフラのコードリポジトリへのアクセスを制限
  - IaCツールにロールベースのアクセス制御を導入
  - インフラコードリポジトリへのアクセスを定期的に監査
- **基礎的なガードレール**
  - インフラの安全な構成と管理については、NIST 800-53ガイドラインに従うこと。
  - クラウドインフラに対して、CSA CCM に概説されているセキュリティ対策を実施。

2.4 MLOpsパイプラインセキュリティ

MLOpsのパイプラインセキュリティにおける重要な領域には、脆弱性のためのソースコードスキャン、攻撃に対するモデルの堅牢性のテスト、各段階でのパイプラインの完全性の検証、自動化スクリプトの監視などがあります。

2.4.1. 脆弱性のためのソースコードスキャン

MLコードの脆弱性のためのソースコードスキャンは、機械学習（ML）アプリケーションのソースコードにセキュリ ティ上の脆弱性やコーディング上の欠陥がないかどうかを、自動化されたツールを使って体系的に調べることです。この作業は、ML システムを危険にさらす可能性のある潜在的なセキュリティ上の弱点を早期に発見し、是正する上で極めて重要です。スキャンでは通常、バッファオーバーフロー、インジェクションの欠陥、安全でないライブラリの使用、意図しない動作やパフォーマンスの問題につながる可能性のあるコーディング手法など、一般的な脆弱性をチェックします。MLのコードを定期的にスキャンすることで、開発者とデータサイエンティストは、コードベースがセキュリティのベストプラクティスと標準に準拠していることを確認し、悪用や侵害のリスクを低減することができます。

このプロアクティブなアプローチは、MLアプリケーションの完全性と信頼性を維持するために不可欠であり、特に機密データを扱ったり、重要なシステムで使用されたりする場合はさらに不可欠です。

- **評価基準**：モデルのトレトレーニングとデプロイに使用されるソースコードのスキャンされた割合と、これらのスキャンの頻度によって、有効性を評価する。
- **RACI**モデル：
    - Responsible: モデル開発チーム
    - Accountable: 最高情報セキュリティ責任者（CISO）
    - Consulted: アプリケーション・セキュリティ・チーム
    - Informed: 開発オペレーション（DevOps）チーム
- **ハイレベルな実装戦略**：ソースコードの脆弱性を定期的にスキャンする自動化ツールを導入し、これらのツールを開発ライフサイクルに統合する。
- **継続的なモニタリングと報告**：コードスキャンの継続的な監視システムを設定し、発見された内容をリアルタイムで報告する。
- **アクセス制御のマッピング**：許可された担当者のみがソースコードとスキャンツールにアクセスし、変更できるようにする。
- **基礎的なガードレール**：NIST 800-53 規格と CSA CCM を指針として活用する。

242. 攻撃に対するモデルの堅牢性のテスト

機械学習（ML）における攻撃に対するモデルの堅牢性のテストは、MLモデルが様々な敵の攻撃や操作にどれだけ耐え、反応できるかを判断するためにMLモデルを評価することを含みます。このテストは、MLモデルにおける潜在的な脆弱性を特定し、それが悪用されて不正な結果を生成したり、システムの誤動作を引き起こしたり、機密情報を漏洩したりする可能性があることを明らかにするために極めて重要です。通常、このような攻撃に対する耐性を評価するために、意図的に細工された入力（敵対的な例）でモデルを厳密に調査すること、異なる脅威シナリオ下でのモデルの動作を分析すること、予期しない入力や悪意のある入力に直面しても性能や精度を維持する能力を検証することが含まれます。堅牢性テストは、特に自律走行車、金融システム、ヘルスケア診断など、堅牢な意思決定が重要なアプリケーションにおいて、MLモデルの信頼性と安全性を確保するのに役立ちます。

- **評価基準**：攻撃に対してテストされたモデルの割合と、これらのテストの頻度によって有効性を測定する。
- **RACI**モデル：
    - Responsible: AI/MLテストチーム
    - Accountable: AI/ML開発責任者
    - Consulted: セキュリティアナリスト
    - Informed: AI/ML開発チーム
- **ハイレベルな実装戦略**：潜在的な攻撃ベクトルを特定し、それを軽減することに重点を置き、モデルのための強固なテストフレームワークを開発する。
- **継続的なモニタリングと報告：**モデルの堅牢性を継続的に評価する仕組みを導入し、利害関係者に定期的に報告する。
- **アクセス制御ののマッピンググ：**テストフレームワークとモデルへのアクセスを適切に制する。
- **基礎的なガードレール**：NIST AI RMF、NIST AI 100-2 E2023
  Adversarial Machine Learning: 攻撃と防御の分類法と用語、およびその他の関連規格。

2.4.3. 各ステージにおけるパイプラインの完全性の検証

機械学習（ML）における各段階でのパイプラインの完全性の検証とは、MLパイプラインの各段階（データ収集、前処理からモデルのトレーニング、評価、デプロイまで）が正しく安全に動作することを保証するプロセスを指します。これには、データの破損、不正アクセス、パイプラインのパフォーマンスとモデルの精度を損なう可能性のあるその他の脆弱性から保護するために、各段階で徹底的なチェックと検証を行うことが含まれます。このような検証には、データの品質と一貫性の検証、安全なデータの取り扱い方法の確保、モデルトレーニングプロセスの信頼性と再現性の評価、デプロイメカニズムが安全で意図したとおりに機能することの確認などが含まれます。このような検証に対する包括的なアプローチは、MLパイプラインの全体的な完全性と有効性を維持するために不可欠であり、特にMLモデルの精度と信頼性が最重要となる複雑な環境や利害の大きい環境では、非常に重要です。

- **評価基準**：MLOpsパイプラインの完全性を綿密にモニタリングすることに重点を置く。これは、検証を受けるステージの割合を調べ、検証プロセスの深さと徹底性を評価することで達成される。その目的は、パイプラインの各ステージが意図したとおりに機能し、確立された標準とベストプラクティスを遵守していることを確認することである。
- **RACI**モデル：
    - Responsible: DevOpsチームは、パイプラインの各ステージを検証する日々のタスクを任されている。彼らは検証プロセスの主要な実行者であり、MLOpsパイプラインの各ステージが徹底的にチェックされ、検証されることを保証する。
    - Accountable: エンジニアリング部門の責任者は、MLOpsパイプラインの全体的な完全性について最終的な責任を負う。この役割には、検証プロセスを監督し、パイプラインが必要な標準と要件を満たしていることを確認することが含まれる。
    - Consulted: 品質保証（QA）チームは相談役であり、検証プロセスに専門家の助言と意見を提供する。QAチームの関与は、検証基準の定義と検証結果のレビューにおいて極めて重要である。
    - Informed: プロジェクトマネージャーは、ステータスと結果について常に知らされる。これにより、プロジェクトのタイムラインや成果物に影響を与える可能性のある問題や変更を確実に把握することができる。
- **ハイレベルな実装戦略**： MLOpsパイプラインの各段階の完全性を検証するためには、体系的なアプローチが不可欠である。この戦略には、データとプロセスの完全性に関する明確な手順と基準を確立することが含まれる。各パイプラインステージに対して具体的な検証テストとチェックを定義し、データの取り込みからモデルのデプロイまで、すべての要素が正しく安全に機能するようにする。
- **継続的なモニタリングと報告**：パイプラインの完全性を検証するために不可欠な要素は、継続的な検証とリアルタイムの報告のためのシステムを導入することである。このシステムは、不一致や異常が発生したときにそれを検知し、問題を修正するための迅速な対応を可能にするものでなければならない。継続的な監視により、パイプラインは常に安全で効率的な状態を保つことができる。
- **アクセス制御のマッピング**：厳格なアクセス管理は、MLOpsパイプラインの各ステージの完全性を維持するために不可欠である。これには、パイプラインのさまざまな部分に対して、誰が、どのような条件で、どの程度の権限でアクセスできるかを定義し、実施することが含まれる。このような管理は、パイプラインの完全性を損なう可能性のある不正アクセスや変更を防止するために不可欠である。
- **基礎的なガードレール**：ベストプラクティスに確実に従うためには、パイプラインの検証プロセスを、NIST Secure Software Development Framework （SSDF）に概説されているような、確立された業界標準やガイドラインに合わせることが重要である。このようなフレームワークに準拠するこ

とで、セキュリティと効率性のベンチマークが提供され、堅牢で信頼性の高い MLOps パイプラインの開発と保守の指針となる。

244. モニタリングオートメーションスクリプト

このタスクには、データの前処理からモデルのデプロイと管理まで、機械学習のライフサイクルのさまざまな段階を自動化するすべてのスクリプトを注意深く監視することが含まれます。

- **評価基準**：自動化スクリプトの監視の有効性は、継続的な監視下にある自動化スクリプトの割合と、監視活動の頻度という2つの主な指標によって定量化される。この評価により、すべてのスクリプトが正確かつ効率的に機能し、潜在的な問題が迅速に特定されて対処されることが保証される。
- **RACI**モデル：
    - Responsible: IT運用チームは、主にMLOpsパイプライン内の自動化スクリプトの日々の監視を担当する。その役割には、スクリプトの実行の監督、パフォーマンスとセキュリティの確保、運用上の問題の特定などが含まれる。
    - Accountable: 最高技術責任者（CTO）は、自動化スクリプトの管理とセキュリ ティに関する全体的な説明責任を負う。CTOは、モニタリング戦略が効果的に実施され、組織の技術目標に合致していることを確認する。
    - Consulted: DevOpsチームは、特にスクリプトのデプロイと運用の効率化において、重要なインプットと専門知識を提供する。パイプライン内で使用されるモニタリングプロセスとツールの強化には、彼らのコンサルテーションが欠かせない。
    - Informed: データサイエンティスト、MLエンジニア、プロジェクトマネージャーなど、MLOpsパイプラインのすべての関係者は、自動化スクリプトのステータスとパフォーマンスについて常に情報を得ることができる。これにより、パイプラインの全ステージにわたって一貫した透明性の高い運用が保証される。MLOpsパイプライン全体で一貫した透明性の高い運用を保証するには、データサイエンティスト、機械学習エンジニア、プロジェクトマネージャーを含むすべての関係者が、自動化スクリプトのステータスとパフォーマンスについて徹底的に情報を得ることが極めて重要である。このプラクティスは、パイプラインの全段階にわたって統一されたアプローチを促進するだけでなく、意思決定が最新かつ正確な情報に基づいて行われることを保証し、AI導入の全体的なセキュリティと効率を高める。
- **ハイレベルな実装戦略**：すべての自動化スクリプトに包括的な監視システムを導入することが 不可欠である。このシステムは、スクリプトのパフォーマ ンスと効率を追跡し、定義された標準とプラクティスに準拠していることを確認す る。 MLOpsパイプラインにシームレスに統合し、自動化スクリプトの動作と出力に関する洞察をリアルタイムで提供する必要がある。
- **継続的なモニタリングと報告**：継続的な監視は、自動化スクリプトの問題を迅速に特定し、対処する上で重要である。監視システムは、リアルタイムのアラートとレポートを生成し、スクリプトのパフォーマンス、エラー、またはセキュリティ上の懸念に関する情報をタイムリーに提供できる必要がある。この継続的なフィードバックループは、MLOps パイプラインの運用の完全性を維持するために不可欠である。
- アクセス制御ののマッピング：自動化スクリプトとパイプライン全体を保護するためには、厳密なアクセス制御が必要である。これには、スクリプトにアクセス、変更、実行できる者を定義することが含まれる。アクセスは役割に応じた要件に基づいて行い、権限を与えられた担当者のみが変更を行えるようにすることで、不正または有害な変更のリスクを低減する。

- **基礎的なガードレール**： CSA CCM のような確立されたフレームワークや、NIST が提供する関連ガイダンスのベストプラクティスを採用する。

2.5 AIモデルガバナンス

AIモデルガバナンスには、モデルリスクアセスメント、ビジネス承認手続き、モデルモニタリング要件、新しいモデルの検証プロセスなど、いくつかの重要な分野が含まれます。

2.5.1. モデルリスクアセスメント

機械学習（ML）におけるモデルリスクアセスメントとは、MLモデルのデプロイと使用に関連する潜在的なリスクを体系的に評価することである。このアセスメントは、モデルの不正確さ、バイアス、または失敗がもたらす可能性のある悪影響を特定し、定量化することを目的としています。主な重点分野には、異なるデータセットやシナリオにわたるモデルの精度と汎化可能性の評価、偏った結果や不公正な結果の可能性の評価、エッジケースや敵対的な条件下でのモデルの動作の理解などが含まれます。モデルリスクアセスメントでは、特に医療、金融、公共安全などの重要なアプリケーションにおいて、モデルが失敗した場合の結果も考慮します。このプロセスは、MLモデルの限界と潜在的な影響を明確に理解することで、MLモデルが責任を持って安全にデプロイされることを保証するために、リスクを特定し緩和するために不可欠です。

- **評価基準**：リスクアセスメントを受けたモデルの割合とその包括性を評価。
- **RACIモデル**：
    - Responsible: リスク管理チーム、データガバナンス委員会
    - Accountable: 最高リスク責任者（CRO）
    - Consulted: AI倫理委員会、法律顧問
    - Informed: データサイエンスチーム
- **ハイレベルな実装戦略**：バイアス、公平性、データプライバシーなど、AIモデルに関連するリスクを評価する枠組みを開発。
- **継続的なモニタリングと報告**：継続的なリスクモニタリングのためのツールを導入し、ソフトウェアサプライチェーンの一部としてリスクを報告するためのプロトコルを確立。
- **アクセス制御のマッピング**：リスクアセスメントツール及びデータへのアクセスが厳格に管理・監視されることを保証。
- **基礎的なガードレール**：NIST AI RMF および NIST 800-53 と整合させ、リスク管理を実施。

2.5.2. ビジネス承認手続き

これは、MLモデルを本番環境にデプロイすることを承認するために組織が従う正式なプロセスとプロトコルを包含します。これらの手順は、MLモデルがビジネス目標に合致し、規制や倫理基準に準拠し、必要なパフォーマ ンスベンチマークを満たしていることを保証します。一般的には、データサイエンティスト、ビジネスアナリスト、リスク管理チーム、場合によっては法務部門やコンプライアンス部門など、さまざまな利害関係者がモデルの有効性、信頼性、潜在的なビジネスインパクトを評価する多段階のレビュープロセスが含まれます。評価される主な側面には、モデルの予測精度、検証データセットでのパフォーマンス、潜在的なバイアスや倫理的問題、データプライバシー法の遵守などがしばしば含まれます。これらの

手順の目的は、MLモデルをデプロイし、ビジネスリスクを最小化し、AI技術の責任ある使用を保証するための、管理され、報告されるアプローチを確立することです。その一例が、OpenAIのPreparedness Frameworkです。

- **評価基準**：デプロイが承認されたAIモデルの割合と、承認ガイドラインの遵守状況を追跡。
- **RACIモデル**：
    - Responsible: プロジェクト管理チーム
    - Accountable: AI最高責任者
    - Consulted: 事業部門リーダー
    - Informed: すべてのAI関係者
- **ハイレベルな実装戦略**：意思決定プロセスに関係する利害関係者を参加させ、AIモデルを承認するための明確な手順を確立。
- **継続的なモニタリングと報告**：定期的に見直す仕組みを持った承認プロセスと決定事項の記録を維持。
- **アクセス制御のマッピング**：承認文書や意思決定ツールへのアクセスを管理。
- **基礎的なガードレール**：NIST SSDF や CSA CCM などのベストプラクティスに従う。

25.3. モデルモニタリング要件

これは、MLモデルが本番環境にデプロイされた後、そのパフォーマンスを追跡・評価する継続的なプロセスを指します。このモニタリングは、時間の経過や様々な条件下でモデルが期待通りに動作し続けることを保証するために極めて重要です。モデルモニタリングの主要な側面は、モデルの予測精度の追跡、モデルの入力や出力のドリフト（データドリフトやコンセプトドリフト）の検出、予測におけるバイアスや不公平の兆候の監視、MLシステムの全体的な健全性とパフォーマンスの監視などです。さらに、モニタリングには、モデルのパフォーマンスにおける重大な変化や異常について、関連する利害関係者に警告を発する仕組みが含まれるべきです。このような継続的な監視は、モデルの劣化、基礎となるデータパターンの変化、または新たなバイアスなどの問題を迅速に特定し、対処するのに役立ち、MLモデルが効果的で公正であり続け、意図した目的に沿ったものであることを保証します。

- **評価基準**：モニタリング活動の頻度と深さに基づいてモデルを評価。
- **RACIモデル**：
    - Responsible: AIオペレーションチーム
    - Accountable: AIオペレーションの責任者
    - Consulted: 品質保証チーム
    - Informed: ビジネスアナリスト
- **ハイレベルな実装戦略**：パフォーマンス、精度、コンプライアンスを追跡する包括的なモデルモニタリングシステムを導入する。最近のMLモニタリング技術には、データやコンセプトのドリフト検出、モデルパフォーマンスの追跡、貢献度による分析（feature attribution analysis）、バイアス検出、リアルタイムアラートなどの機能がある。これらの機能は、リアルタイムの推論データを取得し、ベースラインと比較する SageMaker Model Monitor などの製品に代表される; Google Cloud AI Platform Prediction Monitoringは、モデルの予測とデータドリフトに関する洞察を提供する；、さまざまなAIモニタリングおよび説明可能性プラットフォームは、チームが本番環境でMLモデルを監視、説明、分析し、データドリフト、モデルドリフト、バイアスなどの問題を検出することを可能にする。
- **継続的なモニタリングと報告**：パフォーマンス低下や異常のアラートを持つ継続的なデータ収集や解析のためのシステムを構築。
- **アクセス制御のマッピング**：監視ツールや機微データへのアクセスを制限。
- **基礎的なガードレール**：監視プロトコルについては、NIST 800-53 のガイドラインを活用。

2.5.4 新しいモデルの検証プロセス

このプロセスには、新しいMLモデルを本番環境にデプロイする前に厳密にテストし、検証するための体系的な手順が含まれます。これらのプロセスは、モデルが正確性、信頼性、公平性に関して事前に定義された基準を満たし、不正確な結果や不公正な結果を導きかねない欠陥やバイアスがないことを保証するために設計されています。検証には通常、モデルの性能と汎化可能性を評価するための多様なデータセットに対する広範なテスト、潜在的なバイアスや倫理的問題の検査、敵対的攻撃やデータ異常に対するモデルの堅牢性の評価などが含まれます。さらに、検証プロセスでは、業界標準やベストプラクティスへの準拠を確認するために、モデルの文書や開発手法のレビューを含みます。これらの検証プロセスは、新しいモデルの機能と展開の準備に対する信頼性を確立し、モデルが意図したとおりに機能し、ビジネス目標と倫理ガイドラインに沿った価値を提供することを保証することを目的としています。

- **評価基準**：完全な検証を受けたモデルの割合により、検証プロセスの徹底度を測定。
- **RACIモデル**：
    - Responsible: AI開発チーム
    - Accountable: 最高データ責任者または最高技術責任者
    - Consulted: ITセキュリティチーム
    - Informed: 経営幹部
- **ハイレベルな実施戦略**：正確性、バイアス、セキュリティの脆弱性のテストを含む、新しいモデルを検証するための厳格なプロセスを開発。
- **継続的なモニタリングと報告**：新モデルデプロイ後の継続的な評価のためのプロトコルを確立。
- **アクセス制御のマッピング**：誰が新しいモデルを承認しデプロイできるかを厳密に管理。
- **基礎的なガードレール**：検証のベストプラクティスについては、NIST AI RMF に合わせる。

2.6 セキュアなモデルデプロイ

これには、デプロイプロセスが安全で、管理され、組織の標準に沿ったものであることを保証するためのさまざまな実践が含まれます。主な分野には、デプロイ認可手順、カナリアリリース、ブルーグリーンデプロイ、ロールバック機能、デコミッショニングモデルを含みます。

26.1. カナリアリリース

カナリアリリースとは、新しいMLモデルを本番環境にデプロイする際のリスクを最小化するために使用されるテクニックのことで、より広範囲にデプロイする前に、少数のユーザーに徐々にデプロイします。このアプローチにより、チームは限られた規模ですが、実際のデータやユーザーとのインタラクションがある実環境でモデルのパフォーマンスをテストし、監視することができます。

- **評価基準**：カナリアリリースの有効性を、初期デプロイにおける成功率と問題の検出によって監視。
- **RACIモデル**：
    - Responsible: DevOpsチーム
    - Accountable: AIオペレーション責任者または最高技術責任者
    - Consulted: 品質保証（QA）チーム
    - Informed: 製品管理チーム
- **ハイレベルな実装戦略**：実環境で徐々にモデルをテストするために、デプロイプロセスのステップとしてカナリアリリースを実施。
- **継続的なモニタリングとレポート**：カナリアリリースのリアルタイム監視を設定し、問題を迅速に特定。
- **アクセス制御ののマッピング**：指定されたチームメンバーだけがカナリアリリースを開始し、監視。
- **基礎的なガードレール**：NIST SSDF および NIST 800-53 に従ったデプロイガイドラインに従うこと。

26.2. ブルーグリーンデプロイ

ブルーグリーンデプロイとは、機械学習（ML）モデルを含むソフトウェアデプロイにおける戦略であり、"ブルー "と "グリーン "と呼ばれる2つの同一の本番環境を実行することで、ダウンタイムとリスクを削減します。このアプローチは、新しいモデルのデプロイがアプリケーションのパフォーマンスやユーザーエクスペリエンスに大きな影響を与える可能性があるMLにおいて特に有用です。

- **評価基準**：移行のスムーズさとデプロイ中のダウンタイムによるデプロイ戦略を評価。
- **RACIモデル**：
    - Responsible: ITオペレーションチーム
    - Accountable: 最高技術責任者（CTO）
    - Consulted: DevOpsチーム
    - Informed: エンドユーザー
- **ハイレベルな実装戦略**：ブルーグリーンはデプロイ戦略を採用し、新モデルのデプロイに伴うダウンタイムとリスクを軽減。
- **継続的なモニタリングとレポート**：ブルー環境とグリーン環境のパフォーマンスと問題解決を継続的に監視。
- **アクセス制御のマッピング**：両方の環境へのアクセス制御を管理し、セキュリティと完全性を確保。
- **基礎的なガードレール**：関連する NIST ガイドラインのベストプラクティスを活用。

2.6.4. ロールバック機能

（訳注：項番が2.6.3であるべきであるが、原文に従って2.6.4のままとした。）

機械学習（ML）モデルのコンテキストにおけるロールバック機能とは、新しくデプロイされたモデルが予期せぬ動作やパフォーマンスの低下、その他の問題を引き起こした場合に、MLモデルの以前のバージョンや本番環境のチェックポイントに戻すプロセスを指します。これはデプロイ戦略の重要な側面であり、新しいモデルが期待に応えられなかった場合でも、システムの安定性とパフォーマンスを維持できるようにします。

- **評価基準**：必要に応じてロールバックのスピードと成功率で効果を測定。
- **RACIモデル**：
    - Responsible: デプロイチーム（DevOpsチーム）
    - Accountable: AIオペレーション責任者または最高技術責任者
    - Consulted: ITサポートチーム
    - Informed: ビジネスステークホルダー
- **ハイレベルの実装戦略**：デプロイプロセスに、必要に応じて以前のバージョンに戻すための効率的なロールバック機能が含まれていることを確認。
- **継続的な監視とレポート**：デプロイを監視し、ロールバックを必要とする問題を迅速に検出。
- **アクセス制御のマッピング**：ロールバックツールや手順へのアクセスを制御。
- **基礎的なガードレール**：CSA CCM やその他のセキュリティフレームワークと整合させる。

2.6.5. デコミッショニングモデル

MLモデルの廃棄とは、アクティブな本番環境から機械学習モデルを安全かつ体系的に削除するプロセスを指します。このプロセスは、モデルが古くなったり、より高度なバージョンに置き換わったり、進化するビジネス要件やコンプライアンス基準を満たさなくなったりした場合に不可欠です。廃棄は、MLモデルのライフサイクル管理における重要なステップであり、古くなったモデルがシステムの完全性や効率を危険にさらすことがないようにします。

- **評価基準**：廃棄されるモデルが正しく処理される割合と、廃棄プロトコルの遵守状況によりプロセスを評価。
- **RACIモデル**：
    - Responsible: AIメンテナンスチーム（DevOps）
    - Accountable: データガバナンス責任者または最高技術責任者
    - Consulted: 法務・コンプライアンスチーム
    - Informed: AI開発チーム
- **ハイレベルな実装戦略**：データの安全な取り扱いを確保しつつ、古くなったまたは冗長なAIモデルを安全に廃棄するための明確な手順を策定。
- **継続的なモニタリングと報告**：廃棄措置のプロセスが正しく行われていることを確認するためのモニタリングを実施。
- **アクセス制御のマッピング**：廃棄措置ツールやデータへのアクセスを制限。
- **基礎的なガードレール**： NIST AI RMFおよびその他の関連ガイドラインに従った廃止措置の実施に従うこと。

3. 脆弱性管理

AI脆弱性管理は、AIおよびMLシステムを保護し、安全性、機能性、コンプライアンスを維持するための重要な要素です。このセクションでは、このカテゴリーの主要項目について説明します。

3.1. AI/ML資産インベントリ

AI/ML資産インベントリは、AI/MLランドスケープ内のすべての資産を体系的に記録し、更新します。これには、モデルやデータセットだけでなく、API、アルゴリズム、ライブラリ、ソフトウェアサプライチェーンのAI/MLコンポーネントの作成、トレーニング、展開に使用されるサポートソフトウェアやツールも含まれます。インベントリは、潜在的な脆弱性を特定し、リスクを効果的に管理するために不可欠な、使用中のリソースの明確なビューを提供します。使用されるアセット・インベントリは、MLシステムに応じて、モデルカード、データカード、モデルレジストリの形式をとることができる。どのような資産が存在し、それらがどのように相互接続しているかを理解することは、潜在的な脆弱性を特定するために極めて重要です。

- **評価基準**：AI/ML資産インベントリの有効性は、その包括性と定期的な更新によって評価される。包括的なインベントリは、AI/ML環境のあらゆる側面をカバーし、未確認のコンポーネントを残さない。更新頻度も同様に重要であり、開発された新しいモデル、取得されたデータセット、ソフトウェア環境の変更など、インベントリがAI/MLエコシステムの現状を反映していることを保証する。
- **RACIモデル**：
    - Responsible: IT運用チームは、インベントリの日々の管理と更新を担当する。(またはDevOps）
    - Accountable:最高情報責任者（CIO）または最高技術責任者（CTO）がプロセスを監督し、インベントリが正確に維持され、脆弱性管理に効果的に使用されるようにする。
    - Consulted: AI/ML開発チームは、インベントリを最新かつ適切な状態に保つために必要な洞察や情報を提供する。
    - Informed:経営幹部はAI/ML資産の状態と健全性を常に把握できるため、より高いレベルの意思決定が可能になる。
- **ハイレベルな実装戦略**：AI/ML資産インベントリのレビューと更新の定期的なスケジュールを設けるべきである。これは、AI/ML環境の変更を追跡し、インベントリを更新するよう担当チームに警告する自動化ツールによって促進することができる。このプロセスには、正確性と完全性を確保するための定期的な監査も含まれるべきである。
- **継続的なモニタリングとレポーティング**：リアルタイムの監視システムを導入することで、以下のことが可能になる。
  AI/ML資産の変更を迅速に特定する。これには、モデルの新規導入、既存モデルの更新、データセットの変更、ソフトウェア環境の変更などが含まれる。継続的なモニタリングは、効果的な脆弱性管理に不可欠な最新のインベントリの維持を支援する。
- **アクセス制御のマッピング**：AI/ML資産目録へのアクセスを制限することは、その完全性と機密性を維持するために極めて重要である。アクセスは、インベントリと対話する必要性に応じて、役割ごとに異なるアクセスレベルで、権限を与えられた担当者に制限されるべきである。

- **基盤となるガードレール**：NIST AI RMF（リスク管理フレームワーク）などのフレームワークへの準拠により、AI/ML資産目録が業界のベストプラクティスと規制要件に沿った形で管理されることが保証される。これらのフレームワークは、AI/ML資産の効果的なカタログ化と管理に関するガイドラインを提供し、強固な脆弱性管理戦略に貢献する。

3.2 継続的な脆弱性スキャン

これは、セキュリティ上の弱点を特定するために、すべてのAI/ML資産を体系的かつ継続的に調査することを指します。これには、モデル、データセット、関連インフラ（古いライブラリや安全でないAPI）、AI/ML環境内のその他のコンポーネントのスキャンが含まれます。目的は、悪用される可能性のある脆弱性を検出し、潜在的なセキュリティ問題に先手を打って対処することです。

- **評価基準**：このスキャニング・プロセスの有効性は、AI/ML 資産に対するスキャニングの割合と、スキャニングの頻度という 2 つの主要な指標によって評価される。最適な脆弱性スキャニング・プログラムは、どのコンポーネントもチェックされないまま放置されることがないようにし、アップデートや環境の変化によって生じる可能性のある新たな脆弱性を捕捉するために定期的にスキャニングを実施する。
- **RACIモデル**：
    - Responsible:セキュリティオペレーションチームがスキャンプロセスを実施し、ツールやテクノロジーを活用して徹底的な評価を行う。
    - Accountable:最高情報セキュリティ責任者（CISO）がプロセス全体を監督し、スキャンが効果的に実施され、脆弱性に迅速に対処できるようにする。
    - Consulted: AI/MLチームは、AI/ML資産の特定の要件と構成に関する洞察を提供し、より的を絞ったスキャンを支援する。
    - Informed: IT管理者は、スキャン結果と、より広範なITインフラストラクチャに影響を及ぼす重大な脆弱性について最新情報を得ることができる
- **ハイレベルな実装戦略**：自動化されたスキャンツールの導入は、以下の点で不可欠である。
効率的かつ効果的な脆弱性スキャン。これらのツールは、すべてのAI/ML資産を定期的にスキャンし、新たな脅威の出現に応じて更新するように設定する必要がある。定期的な評価をスケジューリングすることで、AI/ML環境の安全性を長期にわたって維持することができる。
- **継続的なモニタリングと報告**：アラートシステムを確立することは、以下を即座に実行するために極めて重要である。
新たな脆弱性を特定する。このシステムは、検出された脆弱性を関連チームに通知し、迅速な対応と修復を可能にする。
- **アクセス制御のマッピング**：脆弱性スキャンツールおよび結果へのアクセスは、以下のようにすべきである。
厳重に管理される。権限を与えられた担当者のみがスキャンを実施し、詳細な結果にアクセスできるようにし、スキャン中に明らかになった機密情報のセキュリティを確保する。
- **基礎的なガードレール**：NIST 800-53 などの確立されたセキュリティ基準の遵守
脆弱性スキャンプロセスが業界のベストプラクティスに沿ったものであることを保証する。これらの基準は、脆弱性を効果的に特定し、対処するためのガイドラインを提供し、AI/ML システムの全体的なセキュリティ態勢を強化する。

3.3. リスクに基づく 優先順位付け

このプロセスでは、AI/ML資産で見つかった脆弱性を、潜在的な影響と悪用の可能性に基づいて評価し、ランク付けします。このプロセスにより、組織はリソースと労力を最も重要な脆弱性の緩和に集中させることができ、AI/MLシステム全体のリスクを効率的に低減することができます。

- **評価基準**：このアプローチの有効性は、特定された脆弱性の総数と比較して、対処に成功した高リスクの脆弱性の割合によって評価される。高い割合は、最も深刻なリスクの優先順位付けと修復が効果的に行われていることを示す。
- **RACIモデル**：
    - Responsible:リスク管理チームは、リスクレベルに基づいて脆弱性を評価し、優先順位をつけることを任務とする。
    - Accountable:最高情報セキュリティ責任者（CISO）がプロセスを監督し、最も重要な脆弱性が特定され、速やかに対処されるようにする。
    - Consulted:コンプライアンスチームは、特に脆弱性の規制面やコンプライアンス面に関して意見を提供する。
    - Informed:AI/ML開発チームは、各自の資産に影響を及ぼす脆弱性の優先順位と状況について情報を得る。
- **ハイレベルな実装戦略**：包括的なリスク評価の枠組みを構築することは極めて重要である。この枠組みには、機密性、完全性、可用性、悪用の可能性に対する潜在的な影響など、脆弱性の重大性を評価する基準を含めるべきである。
- **継続的なモニタリングと報告**：脆弱性とそのリスクレベルを継続的に監視するシステムの導入は不可欠である。脆弱性のリスク状況を定期的に更新し、報告することで、すべての利害関係者が現在の脅威の状況と改善努力の進捗状況を把握できるようになる。
- **アクセス制御のマッピング**：リスク評価ツール及び脆弱性データへのアクセスは厳重に管理されるべきである。プロセスの完全性を維持するために、権限のある担当者のみが脆弱性の評価、分類、優先順位付けを行えるようにすべきである。
- **基礎的なガードレール**：NIST AI Risk Management Framework(RMF)のようなガイドラインを遵守することで、プロセスが業界のベストプラクティスに合致し、AI/MLシステムのリスクを管理するための構造化されたアプローチが提供される。

リスクベースの優先順位付けは、AIの脆弱性管理において不可欠な要素であり、組織はAI/MLシステムにおける最も差し迫ったセキュリティリスクを軽減するためにリソースを効率的に割り当てることができる。

3.4. Remediation Tracking

これには、AI/MLシステム内で特定された脆弱性に対処し、是正するためのプロセスの継続的な監視と管理が含まれます。脆弱性を緩和するために取られた措置の追跡を含み、脆弱性が速やかに解決されることを保証します。

- **評価基準**：Remediation Trackingの有効性は2つの基準に基づいて評価される：脆弱性の修復に要した時間と、解決した問題の割合である。脆弱性の修復にかかる時間が短く、解決された脆弱性の割合が高いほど、効率的で成功した修復作業であることを示す。脆弱性の修復に要した時間は、上記で定義したリスクに基づく脆弱性の優先順位付けに基づき、組織の脆弱性修復SLAに照らして追跡する必要がある。
- **RACIモデル**：
    - Responsible: IT オペレーションチームは、脆弱性の修復に必要なアクションを実行する。
    - Accountable:最高情報セキュリティ責任者（CISO）は、是正プロセスの有効性に関する全体的な説明責任を負い、脆弱性が迅速に対処されるようにする。
    - Consulted:AI/ML開発チームは、AI/ML資産に関連する脆弱性を是正するための具体的な要件を理解するために意見を提供し、支援する。
    - Informed:エグゼクティブ・リーダーシップは、脆弱性是正の取り組み状況や組織への潜在的な影響について常に情報を得ることができる。
- **ハイレベルな実施戦略**：効率的な修復の追跡には、強固な追跡システムの導入が不可欠である。これらのシステムには、特定された脆弱性の詳細、修復のために取られた措置、責任者、および解決までのスケジュールを記録する必要がある。
- **継続的なモニタリングと報告**：進捗状況の更新や完了状況など、修復活動の詳細な記録を維持することが重要である。継続的な監視により、脆弱性が正常に解決されるまで、積極的に追跡・管理されるようにする。
- **アクセス制御のマッピング**：改善活動に関連する文書へのアクセスは、不正アクセスや改ざんを防止するために、安全に管理されなければならない。これにより、改善プロセスの完全性を保護する。
- **基礎的なガードレール**：NIST 800-53 などの確立されたサイバーセキュリティ標準を参照することで、修正追跡プロセスが業界のベストプラクティスに合致し、脆弱性の修正を管理および文書化するための構造化されたアプローチが提供される。

3.5. 例外処理

例外処理とは、確立されたセキュリティプロトコルや手順から逸脱したり、例外が発生したりする状況を効果的に管理するプロセスです。このような例外は、標準的なセキュリティ慣行からの逸脱を必要とする独自の状況、運用上のニーズ、レガシーシステム、その他の要因によって発生する可能性があります。

- **評価基準**：例外処理の有効性は、処理された例外の数と解決の全体的な有効性に基づいて評価される。適切に管理された例外処理プロセスでは、例外の発生件数を最小限に抑え、例外が発生した場合は、代償的な管理策を使用してセキュリティ上の懸念に対処しながら、適切に対処するようにしなければならない。
- **RACIモデル**：
    - Responsible: セキュリティチームは、セキュリティ上の例外を管理し、発生した例外に対処する。

    - Accountable: 最高情報セキュリティ責任者（CISO）は、例外処理プロセスの有効性に関する全体的な説明責任を負い、例外がセキュリティポリシーと規制に沿った形で管理されるようにする。
    - Consulted: 法務およびコンプライアンスチームは、例外が法律および規制要件の範囲内で管理されるよう、指導および助言を行う。
    - Informed:経営陣は、例外とその解決策について知らされ、透明性と業務全体との整合性を確保する。
- **ハイレベルな実装戦略**：例外処理は、例外の発生状況、例外に対処するために取られた措置、および実施された是正措置を含め、十分に文書化されるべきである。これらのプロトコルは、セキュリティが最優先事項であり続けることを確保しつつ、例外を特定し、評価し、対処するためのプロセスを定義すべきである。
- **継続的な監視と報告**：例外処理プロセスおよび関連文書へのアクセスは、権限を与えられた担当者のみに制限されるべきである。継続的なモニタリングと報告により、例外が長期にわたって効果的に管理されるようにする。
- **アクセス制御のマッピング**：例外処理プロセスおよび関連文書へのアクセスは、権限を与えられた担当者のみに制限されるべきである。これにより、例外の取り扱いが安全で、確立されたプロトコルに準拠していることが保証される。
- **基礎的なガードレール**：クラウドなどの業界標準との整合性の確保
  セキュリティアライアンス（CSA）のクラウドコントロールマトリックス（CCM）は、例外処理のベストプラクティスの確立を支援し、認められたガイドラインに従って例外が管理されることを保証する。

例外処理はAI脆弱性管理の重要な要素であり、組織は全体的なセキュリティとコンプライアンス基準を維持しながら、固有の状況に効果的に対応することができます。

3.6. メトリクスの報告

これは、AIの脆弱性管理の取り組みの有効性を評価し、定量化するために使用される具体的な測定と主要業績評価指標（KPI）のことを指します。これらの指標は、AI/MLシステム内のセキュリティ状態に関する貴重な洞察を提供します。

- **評価基準**：報告の正確性と適時性は、レポーティング・メトリクスの有効性を評価する上で不可欠である。正確で最新のレポーティングは、以下を確実にする。
  意思決定者は、十分な情報に基づいたセキュリティ上の意思決定を行うために、信頼できる情報にアクセスすることができる。
- **RACIモデル**：
    - Responsible：報告チームは、脆弱性管理メトリクスの収集、分析、提示に責任を負う。
    - Accountable：最高情報セキュリティ責任者（CISO）は、報告プロセスの全体的な有効性に責任を負い、評価指標がセキュリティ目標に合致していることを確認する。
    - Consulted：AI/ML と IT 部門は、AI/ML 環境のセキュリティ態勢を正確に反映した指標となるよう、インプットとコンテキストを提供する。
    - Informed：経営幹部は、報告指標から得られる結果と洞察について知らされ、戦略的な意思決定を行うことができる。

- **ハイレベルな実装戦略**：標準化された報告手順を作成することが重要である。これらの手順は、一貫性と正確性を確保するために、測定基準の収集、分析、提示方法を定義する必要がある。
- **継続的なモニタリングと報告**：定期的な更新とレポートの配布
  AIの脆弱性管理の現状について、すべての利害関係者に情報を提供し続けることが不可欠である。この継続的な報告により、セキュリティ問題が迅速に特定され、対処されることが保証される。
- **アクセス制御のマッピング**：レポーティングツールおよびデータへのアクセスは、以下を防ぐために制御されるべきである。
  不正アクセスやメトリクスの改ざん。アクセスを制限することで、報告プロセスの完全性を保護する。
- **基礎的なガードレール**：以下のような認知されたフレームワークのベストプラクティスに従う。
  米国国立標準技術研究所（NIST）のセキュアソフトウェア開発フレームワーク（SSDF）とクラウドセキュリティアライアンス（CSA）のクラウドコントロールマトリックス（CCM）は、効果的な報告メトリクスの開発と管理のための業界標準ガイドラインを確立するのに役立つ。

結論

本ホワイトペーパーでは、AIやMLシステムを開発・導入する際に組織が守るべき中核的なセキュリティ責任について検討しました。データセキュリティ、モデルセキュリティ、脆弱性管理に焦点を当てることで、AIシステムのライフサイクル全体を通じてセキュリティ、プライバシー、完全性を確保するための包括的なフレームワークを概説しました。

データセキュリティとプライバシーの分野では、データの真正性、匿名化、仮名化、データの最小化、アクセス制御、安全な保管と送信の重要性を強調しました。これらの対策は、機密情報を保護し、データ保護規制へのコンプライアンスを維持するために不可欠です。

モデルのセキュリティに関しては、アクセス制御、セキュアなランタイム環境、脆弱性とパッチの管理、MLOpsパイプラインのセキュリティ、AIモデルのガバナンス、セキュアなモデル展開の重要性について議論しました。強固なセキュリティ管理とガバナンスプロセスを導入することで、組織はAIモデルに関連するリスクを軽減し、信頼性の高い運用を確保することができます。

脆弱性管理は、AIセキュリティのもう一つの重要な側面です。私たちは、AI/ML資産のインベントリの維持、継続的な脆弱性スキャンの実施、リスクの優先順位付け、改善努力の追跡、例外処理、報告指標の確立の必要性を強調しました。これらの実践により、組織は脆弱性をプロアクティブに特定して対処することができ、セキュリティ侵害の可能性を最小限に抑え、AIシステムの継続的なセキュリティを確保することができます。

文書全体を通じて、定量化可能な評価基準、役割定義のためのRACIモデル、ハイレベルな実装戦略、継続的な監視と報告の仕組み、アクセス制御のマッピング、NIST AI RMF、NIST SSDF、NIST 800-53、CSA CCMなどの業界のベストプラクティスや標準に基づく基礎的なガードレールの順守を使用して、各責任を分析しました。

本ホワイトペーパーに記載されている推奨事項とベストプラクティスを採用することで、企業は安全で責任あるAIの開発と導入のための強固な基盤を確立することができます。しかし、AIのセキュリティは継続的なプロセスであり、テクノロジーや脅威の進化に合わせて継続的な監視、適応、改善が必要であることを認識することが不可欠です。

組織がAI導入の複雑さを乗り越えるには、経営陣、技術チーム、ガバナンス組織、エンドユーザーを含むすべての関係者の間で、セキュリティとコラボレーションの文化を醸成することが極めて重要です。本ホワイトペーパーで取り上げた原則と実践を遵守し、協力することで、組織は、関係者全員のセキュリティ、プライバシー、信頼を確保しながら、AIの変革の可能性を引き出すことができます。

略語

AI	Artificial Intelligence
AI RMF	Artificial Intelligence Risk Management Framework
AIMS	AI Management System
AIOps	Artificial Intelligence for IT Operations
API	Application Programming Interface
AWS	Amazon Web Services
CAIO	Chief AI Officer
CCM	Cloud Controls Matrix
CDO	Chief Data Officer
CEO	Chief Executive Officer
CFO	Chief Financial Officer
CI/CD	Continuous Integration/Continuous Deployment
CIO	Chief Information Officer
CIS	Center for Internet Security
CISO	Chief Information Security Officer
CNAPP	Cloud Native Application Protection Platform
COO	Chief Operating Officer
CPO	Chief Privacy Officer
CSA	Cloud Security Alliance
CTO	Chief Technology Officer
CVE	Common Vulnerabilities and Exposures
CVSS	Common Vulnerability Scoring System
DAST	Dynamic Application Security Testing
DataOps	Data Operations
DDoS	Distributed Denial of Service
DevSecOps	Development, Security, and Operations
DISA	Defense Information Systems Agency
DoS	Denial of Service
ENISA	European Union Agency for Cybersecurity
GDPR	General Data Protection Regulation
HSM	Hardware Security Module
IaC	Infrastructure as Code
IaaS	Infrastructure as a Service
IAM	Identity and Access Management
IDPS	Intrusion Detection and Prevention System
IDS	Intrusion Detection System
IEC	International Electrotechnical Commission
IPS	Intrusion Prevention System
ISM	Information Security Manager
ISMS	Information Security Management System
ISO	International Organization for Standardization
ISS	Information Systems Security
ISSO	Information Systems Security Officer

K8s	Kubernetes
KPI	Key Performance Indicator
LLM	Large Language Model
MFA	Multifactor Authentication
MITRE ATT&CK	MITRE Adversarial Tactics, Techniques, and Common Knowledge
ML	Machine Learning
MLOps	Machine Learning Operations
NIST	National Institute of Standards and Technology
OS	Operating System
OWASP	Open Web Application Security Project
PaaS	Platform as a Service
PIMS	Privacy Information Management System
PoLP	Principle of Least Privilege
QA	Quality Assurance
RACI	Responsible, Accountable, Consulted, Informed
RBAC	Role-Based Access Control
SaaS	Software as a Service
SAST	Static Application Security Testing
SDLC	Software Development Life Cycle
SLA	Service Level Agreement
SSDF	Secure Software Development Framework
STIGs	Security Technical Implementation Guides
TEE	Trusted Execution Environment
TLS	Transport Layer Security
VPN	Virtual Private Network
WAF	Web Application Firewall