

エージェントによる実行を統制する

セキュアなエージェント企業のための
アーキテクチャブループリント

エージェントは
どこにいるのか

エージェントは
何ができるのか

エージェントは
何をしているのか？

どのように
対応すべきか？

Blueprint Allianceの創設メンバー執筆

CISO、CIO、CTO、エンタープライズセキュリティアーキテクトのための
業界横断リファレンスアーキテクチャ

目次

1	エグゼクティブサマリー
2	はじめに
3	Blueprintの原則
4	「Blueprint」のリファレンスアーキテクチャ
5	グローバルシステムインテグレーター（GSI）との連携による運用
6	まとめ：AIセキュリティの未来の形成
7	Blueprintの実践
8	利用開始に向けたベストプラクティス

エグゼクティブサマリー

AIエージェントの登場により、従来のエンタープライズセキュリティの境界が脅かされています。実験的なLLMチャットボットは、わずか2年足らずのうちに自律的なマルチステップのエージェント型ワークフローへと進化しました。かつてソフトウェアは、指示されたことを実行していました。しかし今では、自ら判断して行動し、独自に接続を行う存在となっています。そうした自律性はビジネスに大きな価値をもたらす一方で、多くのセキュリティチームが対応できないガバナンス上の課題をもたらしているのが現状です。

[Gartnerの予測](#)によると、Fortune 500に選出されているグローバルエンタープライズにおいて、エージェントの平均利用数は2025年の15未満から2028年には15万以上へと急増します。これにより、エージェントの無秩序な増加やITの複雑化、管理上の課題が深刻化するとみられています。このようなエージェントは、単にデータを表示するだけではありません。推論、ルーティング、ダウンストリームAPIの呼び出し、サブエージェントの生成を行い、多数のエンタープライズシステムにまたがってマルチステップの操作を実行します。

統一されたアーキテクチャ標準がない場合、組織は**アイデンティティ、セキュリティ、ガバナンスにおけるギャップの拡大という課題に直面します**。つまり、シャドーエージェントが急増し、信頼境界を越えてアクセス権が引き継がれ、実行が制御不能になってしまいます。Blueprint Allianceのリファレンスアーキテクチャは、オープンかつマルチなベンダーフレームワークであり、オープンコミュニティで利用できるよう公開されています。このフレームワークでは、セキュリティの基盤をネットワーク境界から、統合されたゼロトラストのエージェント型コントロールプレーンへと移行することで、テクノロジーリーダーが制御力を失うことなくエージェント型デプロイメントを拡張するための具体的な方法が提供されます。

はじめに

長年の間、エンタープライズセキュリティは、人間のアイデンティティ (IAM、MFA、SSOによって管理されるもの) とマシンのアイデンティティ (静的なサービスアカウント、APIキー、クラウドベースのコンピューティングサービスによって管理されるもの) を明確に区別しながら行ってきました。自律型AIエージェントは、この二分法では捉えきれなくなります。AIエージェントが人間の従業員に代わって行動し、アクセス特権を継承し、ベクトルデータベースにクエリを実行し、金融取引を開始する場合、そのAIエージェントは、明確なアイデンティティ、アクセス、ガバナンスの制御を必要とするアクティブなセキュリティプリンシパルおよび運用アクターとなります。

セキュリティは、モデル層やネットワーク層だけに後付けできるものではありません。可視性、接続性、実行を継続的に統制するエージェント型コントロールプレーンを基盤として構築する必要があります。

従来のソフトウェア (決定論的)	エージェント型AI (非決定性)
明示的で段階的なコードパスを実行します。	人間の高レベルの意図を受け取り、複数ステップの実行計画を策定します。
静的で事前定義された権限とサービスアカウントに依存します。	トークンボルト、認証情報ブローカー、委任セッションを活用して、APIとアプリケーションを動的に選択します。
ハードコードされた統合パスに厳密に制限されます。	自律的に動作し、モデルコンテキストプロトコル (MCP) を介してツールを利用し、サブエージェントを生成します。
認証済みユーザーセッションの範囲内で動作します。	人間の委任者がログオフした後も多くの場合、長時間にわたって非同期で稼働します。

ガバナンスのギャップ：

すべての取締役会が抱える5つの問い

AIエージェントは、ほとんどのセキュリティチームが追跡しきれないほど急速に増加しています。[Gartnerの予測](#)によると、Fortune 500に選出されているグローバルエンタープライズにおいて、エージェントの平均利用数は2025年の15未満から2028年には15万以上へと急増します。これにより、エージェントの無秩序な増加やIT環境の複雑化、管理上の課題が深刻化するとみられています。AIエージェントに適切なガバナンスを導入できていると考える組織は、13%にすぎません。こうしたギャップがあるため、サイバーセキュリティおよびコンプライアンス委員会は、次の5つの問いに立ち返って答える必要があります。

1. AIエージェントはどこに存在し、誰が所有しているのか？
2. エージェントが実行できる操作をどのように制限し、タスクから逸脱させずに運用するのか？
3. 自律的な意思決定について、事後に監査し、経緯を説明できるのか？
4. エージェントのサプライチェーン全体におけるデータ漏洩およびサードパーティに起因するリスクには、どのようなものがあるのか？
5. エージェントが制御不能になった場合に、検証済みのキルスイッチとロールバック計画は用意されているのか？

これら5つの問いは、セクション3で解説する「Blueprint」のリファレンスアーキテクチャの4つの柱に対応しています。「ディスカバリーとアイデンティティ登録」が問い1に、「アクセス管理」が問い2に、「ランタイム認可とモニタリング」が問い3と4に、そして「アクティブな封じ込め」が問い5に答える形になっています。CISOは、単なる保証ではなく証拠に基づいてこれらの5つの問いに答えることで、規制上・運用上のリスクに見合った制御レベルを維持しつつ、エージェントの導入を安全に加速できます。これができないと、ますます深刻化・増加するセキュリティインシデントに対して、脆弱性を抱えることになるでしょう。

これらの問いは、画一的なチェックリストではなく、リスクを評価するための診断ツールとして機能します。セキュリティ要件は当然ながら運用範囲によって異なります。読み取り専用の生産性アシスタントに、金融取引を実行するエージェントほどの封じ込め制御は必要としません。実際に該当するリスクを診断することで、セキュリティチームは現実のリスク状況に応じて制御できるようになります。これにより、組織は低リスクの導入に過剰な設計を行ったり、数か月におよぶ承認遅延に陥ったりすることを回避しつつ、高リスクのエージェントに適切な運用上の境界を確実に設定できるのです。

Blueprintの原則

Blueprint Allianceのフレームワークは、ゼロトラストをエージェントにまで拡張します。エージェントをコードや認証情報、あるいは人間のアイデンティティスタックに後付けされた存在ではなく、人間同様の第一級のアイデンティティとして取り扱います。これら6つの原則は、以下のすべてを支える土台となります。

ホストされているすべてのエージェントは、後付けではなく、個別のセキュリティアイデンティティとして扱います。 エージェントには、従業員やサービスアカウントと同様の厳格さでプロビジョニング、認証、デプロビジョニングを行います。パイロットや社内ツールなど、企業でホストされているすべてのエージェントは、そのリスクと運用の範囲に応じた制御レベルで、適切なガバナンスフレームワーク内で運用されるべきです。ローカルランタイムは、正式なアイデンティティディレクトリへの登録ではなく、専用のエンドポイント封じ込めの下で動作します。

アクセス権は永続的ではなく、タスクごとに範囲を設定します。 タスク単位で範囲設定された動的な権限付与が理想的な目標である一方、実際の実装では、オーバーヘッドとのバランスを取るために、有効期限が短いタスク権限付与と回復力のあるベースライン権限を組み合わせます。

委任は、エンドツーエンドで追跡可能なものとしします。 人間がエージェントに委任し、そのエージェントがサブエージェントを生成した場合、説明責任のチェーンは各ホップで維持されなければなりません。すべてのアクションは、それを認可した人間またはプロセスまで遡って追跡する必要があります。エンドツーエンドのマルチホップ委任は、現在のツールでは依然として目標の段階ですが、このBlueprintでは、標準が成熟していく中で参照すべき方向性を定義しています。

エージェントランタイムは、プロビジョニングされるだけでなく、分離・監視されるものです。 セットアップ時に行われるアクセス判断は必要ですが、それだけでは十分ではありません。このBlueprintでは、付与された権限だけでなく、実行全体にわたる継続的な観測を通じて行動のベースラインを確立することが求められます。

封じ込めは即時で、かつ可逆的なものとしします。 ベンダーやプラットフォームを問わず、すべてのエージェントには、リアルタイムでエージェントを一時停止、隔離、または終了できる、対象を絞ったキルスイッチが必要であり、明確な復元できる方法が備わっていなければなりません。継続的な振る舞い監視と構造的な実行境界を組み合わせることで、エージェントがそのベースラインから逸脱した瞬間に影響範囲を封じ込めて、「認可済み」と「安全」とのギャップを解消することができます。

ガバナンスは、AIの速度に合わせて適応させます。 静的なセキュリティポリシーは、急速に進化するエージェント型機能に対応できず、ポリシーだけでは予測できない事態をガバナンスで補足する必要があります。このBlueprintは、新たなエージェントの動作に合わせてスケールする適応型アーキテクチャとして機能します。静的な権限上限と動的なリアルタイムの範囲設定を組み合わせることで、組織は厳格なガードレールを維持しながら、コンテキストに応じてアクセスを適応できるようになります。

これらの6つの原則によって、ゼロトラストが新たなタイプのアイデンティティ、つまり、企業全体で常時、持続的かつ自律的に稼働するAIエージェントにまで拡張されることになります。正しく実装すれば、アイデンティティ、セキュリティ、ガバナンスのギャップは解消されます。

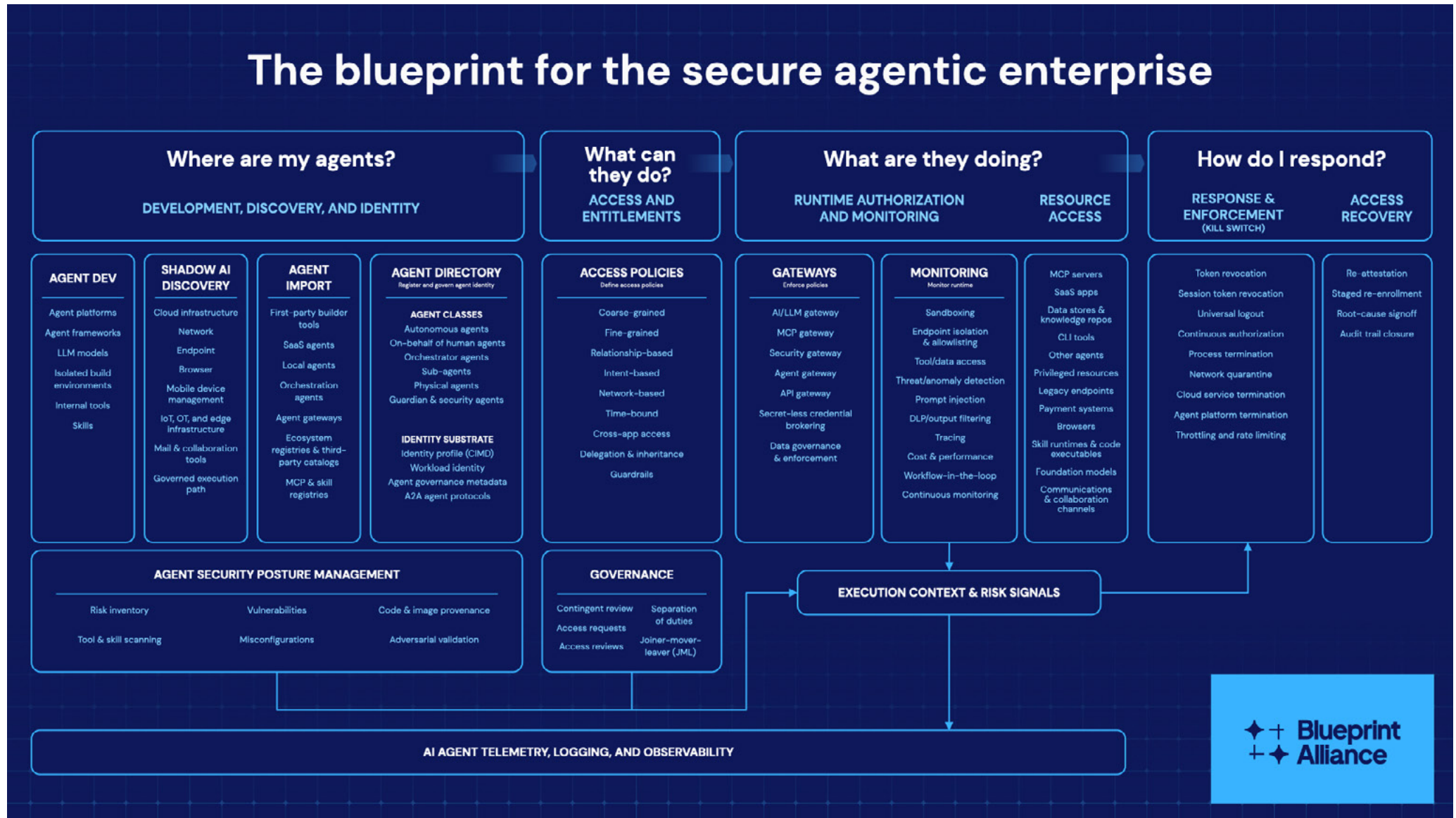
「Blueprint」のリファレンスアーキテクチャ

「Blueprint」のリファレンスアーキテクチャでは、自律的な実行を統制するために不可欠な4つの運用上の問いを軸として、エージェント型企業のセキュリティを構築しています。それぞれの問いは、リファレンスアーキテクチャの主要な柱に対応しています。

1. エージェントはどこにいるのか
2. エージェントは何ができるのか
3. エージェントは何をしているのか？
4. どのように対応すべきか？

すべての問いに共通するのは、実行時のテレメトリ、ロギング、可観測性によって継続的に供給される実行コンテキストとリスクシグナルの強固な基盤が必要です。これらが組み合わさることで、システムがリスクパターンを把握し、連携して応答できる「結合組織」の役割を果たします。

「Blueprint」のリファレンスアーキテクチャ



3.1 第1の柱:

エージェントはどこにいるのか？ (開発、検出、アイデンティティ)

エージェント型企業のセキュリティを確保する上で根本的な課題となるのは、可視性です。存在を把握していないものを、組織は統制・保護できません。この柱では、企業内のあらゆる環境で包括的な追跡機能確立し、承認済みか管理対象外かを問わず、企業インフラ間を横断するあらゆるAIエージェントを特定し、カタログ化・評価します。最終的に、検証済みのエージェントを説明責任を担う所有者を持つ統制対象アイデンティティとして登録します。これにより、第2の柱においてアクセスポリシーと結び付けるための権威インベントリが作成されます。

3.1.1 AIエージェントの検出

AIエージェントの検出は、企業におけるエージェントの稼働領域を継続的に検知する仕組みとして機能します。社内の開発パイプラインから外部のサードパーティ統合まで、AIエージェントのライフサイクル全体が明確になり、稼働中のモデルや自律的なプロセスが見えない場所で動作しないようにします。

3.1.1.1 エージェントの開発

このコンポーネントは、社内の開発環境に焦点を当て、独自エージェントの積極的な構築・学習・調整が対象となります。開発パイプラインを保護することで、エージェントが本番データにアクセスする前に、セキュリティバイデザインを確実に担保できます。

- **エージェントプラットフォーム:** エージェントがホスト、トレーニング、体系的に導入される、企業が管理するオーケストレーション環境です。
- **エージェントフレームワーク:** 開発者がエージェント型のロジックを構築するために利用する、基盤となるソフトウェアライブラリおよびアーキテクチャ (LangChain、LlamaIndex、AutoGenなど) です。
- **LLMモデル:** エージェントを駆動する認知エンジンとして機能する、基盤となりファインチューニング済みの大規模言語モデルです。
- **分離されたビルド環境と実行環境:** テストおよび評価中にエージェントコードを実行する、コンピューティングから分離されたセキュアなCI/CD境界です。これにより、セキュリティチームは静的なレジストリ宣言のみに頼るのではなく、一時的なテスト実行を検出できます。
- **その他の社内ツール:** 開発パイプラインに統合されたカスタムコードリポジトリ、コードアシスタント、社内モデルレジストリです。
- **スキル:** エージェントの機能を拡張するために構築された、定義済みの手続き型モジュール、コードスニペット、またはツールです。

3.1.1.2 エージェントのインポート

現代のエンタープライズが単一のインテリジェンスソースのみに頼ることはほとんどありません。代わりに、急速に変化するエージェント型機能への備えとして、オリジン、フレームワーク、コンピューティング境界が異なる多様なエージェントアーキテクチャを取り込んでいます。エージェントのインポートは、こうした様々なクラスのエージェントが企業エコシステムに導入されるプロセスを監視・統制するもので、その基盤となる導入モデルに基づいて組織全体でデータがどのようにフローするかに重点を置いています。

インポートするエージェントについては、「エージェントタイプ」「カテゴリー」「チャンネル」の3つの独立した問いがあります。通常、1つのエージェントはこれら3つの問いすべてに回答を持っています。エージェントゲートウェイ経由で取り込まれ、社内の人事ワークフロー(B2E)で使用されるSaaSエージェント(ベンダー製)はその一例です。

エージェントタイプ：誰が構築したのか？

- **ファーストパーティエージェント**：生のコード(PythonやLangChainなど)で構築され、汎用コンピューティング(Kubernetesやサーバーレス関数など)で実行される完全カスタム実装から、ランタイム、メモリ、ワークフローのオーケストレーションをネイティブに処理するプロバイダー固有のエージェント型フレームワークを使用して構築されたエージェントまで、社内で開発されたAI機能です。
- **SaaSエージェント**：エンタープライズソフトウェアエコシステム内で完全に独立したエンティティとして動作する、目的特化型のサードパーティ製AI製品です。
- **ローカルエージェント**：従業員のローカルデバイスに直接インストールされるクライアントアーキテクチャです。推論の各ステップがリモートのクラウドLLMプロバイダーに送信されます。人間の意図とツールの実行とに存在する非決定論的なレイヤーとして動作し、推論が呼び出されるたびにデータが組織の境界を越えてデータが送受信されます。
- **オーケストレーションエージェント**：視覚的なオーケストレーションプラットフォーム上に構築されたエージェントで、様々なエンタープライズAPIを「つなぎ合わせる」役割を果たします。

エージェントカテゴリ：誰のためのものか？

- **B2E（従業員向け）エージェント**：社内従業員向けのAIエージェント。企業内部のアイデンティティ管理やロールベースのアクセス制御環境において、企業ワークフロー、IT/HR運用、従業員の日常的な生産性を自動化するために導入されます。
- **B2Bエージェント**：企業の境界を超えて稼働し、企業間のワークフロー、サプライチェーン取引、クロステナントのAPI統合を自動化する、ビジネス向けのAIエージェントです。
- **B2Cエージェント**：サービス、コマース、またはサポートのために外部エンドユーザーと直接やり取りするために導入される、消費者向けのAIエージェントです。パブリックプロンプトインジェクション、ブランドリスク、消費者のPII漏洩に対する厳格な防御が求められます。
- **B2B2Cエージェント**：企業クライアントを通じてそのパートナーの最終消費者にサービスを提供するために提供される、ホワイトラベルまたはベンダー提供のAIエージェントです。多層分離と推移的な信頼境界の下で稼働します。
- **パーソナルアシスタントエージェント**：個人とエンタープライズの両方のコンテキストで個人に代わって稼働する、消費者グレードのユーザー向けのAIアシスタントです。個人データと企業データへのアクセスを区別する境界制御が求められます。

取り込みチャンネル：どのように検出されますか？

- **エージェントゲートウェイ**：外部またはインポートされたこれらのエージェントタイプが、社内エンタープライズAPIとやり取りするための取り込みポイントおよびプロキシです。
- **エコシステムのレジストリとサードパーティのカタログ**：外部のベンダーネイティブなレジストリからエージェントの定義を継続的に監視、取得、マッピングする自動化されたコネクターパイプラインとAPIアダプターです。エージェントが、介入なしで自動的にカタログ化されます。
- **MCPおよびスキルレジストリ**：エージェントが、パブリックまたはベンダーがホストするMCPおよびスキルレジストリから、ツールスキーマ、スキル、コンテキストの定義を取得するための取り込みポイントです。レジストリ自体がインジェクションベクトルであり、悪意のあるエントリや偽装されたエントリが自動的に取得・消費される可能性があるため、レジストリから取得したすべてのツールやスキルは、エージェントが呼び出す前に、信頼できない入力として扱われます。セクション3.1.2で扱うサプライチェーンスキャンの対象となります。

3.1.1.3 シャドー AIの検出

従業員はワークフロー効率化のために、承認されていないAIツールを自然と利用するようになり、コンプライアンス違反、データ窃取、運用セキュリティのリスクをもたらします。シャドー AIの検出は、企業の境界を動的にスキャンして、未認可エージェントの展開や利用を検知します。ディスカバリー機能と監視機能は、適用されるプライバシー、雇用、労働、労使協議会、データ保護の要件に従って実装する必要があります。その際、適切な透明性、比例性、データ最小化、保持、アクセス制御を確保しなければなりません。

- **クラウドインフラストラクチャ:** アクティブなネットワーク/API検査によって検出された、クラウドコンピューティングインスタンス、サーバーレス環境、および管理対象外のAPIエンドポイントです。
- **ネットワーク:** リアルタイムのトラフィック分析とエグレス監視により、外部のLLMプロバイダーやAIエンドポイントへの未認識のAPI呼び出しを特定します。
- **エンドポイント:** デバイスレベルの監視により、AIバイナリ、スクリプト、またはローカルモデルの重みに対する不正な実行を追跡します。
- **ブラウザ:** 拡張機能の追跡とブラウザセッションの監視により、ユーザーが承認されていないWeb ベースのAIプラットフォームやチャットシステムにログインしたことを検知します。

- **モバイルデバイス管理 (MDM) とエンドポイント管理:** デバイスレベルの継続的な監査により、未承認のローカルAI実行ファイル、ローカルモデルの重み、CLIエージェントツール、IDE拡張機能を検出するとともに、ベースラインのセキュリティ態勢を徹底し、オペレーティングシステムレイヤーで管理対象外のエージェントランタイムをブロックします。
- **IoT、OT、エッジインフラストラクチャ:** 従来のエンドポイントソフトウェアを導入できない製造装置、PLC、組み込みハードウェアに対し、小規模言語モデル (SLM) や自律型エージェントを検出する産業用セキュリティ監視およびネットワークプロトコル検査機能です。
- **メールとコラボレーションのフロー:** メッセージ、チャット、カレンダーのトラフィックを分析し、アクティビティがエンドポイントやネットワークのエグレスログに表示される前に、ユーザーに代わってメッセージの送信、自動返信、カレンダーの管理を行う未承認のエージェントを検出します。
- **検出面としての統制された実行パス:** 承認されたすべてのエージェントが、一元管理されたコントロールプレーンを介して実行するアーキテクチャ上の境界です。この境界外で稼働するエージェント型プロセスは、環境固有の個別の検出ルールを必要とせずに、シャドー AIのシグナルとして自動的にフラグ付けされます。

3.1.2 AIエージェントのセキュリティ態勢管理 (AIASPM)

検出されたエージェントは、継続的に評価する必要があります。AIASPMは、エージェントのインベントリ全体にわたってリスクプロファイル、構成の健全性、既知のソフトウェアの脆弱性を評価する診断レイヤーを提供します。

- **リスクインベントリ**: 検出されたエージェントにおいて、ライフサイクル状態に関係なく、セキュリティ態勢、脆弱性、ソフトウェアの依存関係、設定ミスを追跡する、一元化された動的な台帳です。IAMに特化したエージェントディレクトリを補完する、セキュリティリスクの視点を提供します。
- **脆弱性**: 標準的なソフトウェア脆弱性管理をエージェント型の依存関係とモデルのエンドポイントにまで拡張し、エージェントのコードベース、サプライチェーン、基盤モデル全体のCVEを網羅します。これらを標準化されたフレームワークにマッピングすることで、一貫したリスクスコアリングを実現します。
- **コードとイメージの来歴**: 本番環境での実行前に、エージェントコード、コンテナイメージ、依存関係、モデルの重みのオリジン、ビルドの完全性、署名 (SLSAフレームワークの構成証明など) を暗号技術によって検証します。不要なパッケージを削除し、最新のCVEデータに照らして再構築された、事前証明済みの最小限のベースイメージを利用することで、サプライチェーンの攻撃対象領域を縮小し、ビルド時の来歴検証を簡素化します。

- **ツールとスキルのサプライチェーンスキャン**: ローカルエージェントの指示セット、実行スキル、リモートのModel Context Protocol (MCP) ツールスキーマ (MCPおよびスキルレジストリから取得したエントリを含む) を静的および動的に自動検査します。登録前に、埋め込まれた悪意のあるコード、隠蔽された操作、不正なエグレスパスを検出します。
- **設定ミス**: 継続的なスキャンにより、脆弱なパラメーター設定、安全でないシステムプロンプト、ハードコードされた有効期間の長い認証情報、なりすましのベクトル、エージェントの初回レビュー通過後に生じた構成ドリフトを検出します。
- **敵対的検証**: 使い捨ての分離された実行境界内で、既知のプロンプトインジェクション、ジェイルブレイク、データ窃取パスに対してエージェントとそのガードレールの敵対的テストを継続的かつ自動的に実施し、本番環境に影響を与えるリスクを冒すことなく、現実世界での悪用可能性を安全に評価します。

3.1.3 エージェントディレクトリ（エージェントアイデンティティの登録と統制）

エージェントのインポート（セクション3.1.1.2）では、検出中に来歴とメタデータを取得しますが、エージェントディレクトリは、検証済みのエンタープライズホスト型エージェントすべての暗号化認証情報、所有権、ポリシーエンタイトルメントを正式に登録・維持する、信頼できるアイデンティティリポジトリです。登録は、検出されたエージェントと管理されたエージェントを分ける境界線であり、第2の柱におけるすべてのアクセス決定の前提条件となります。

ホストされているすべてのエージェントと、個別の権限を持つすべてのサブエージェントは、エンタープライズシステムへのアクセスを要求する前に、検証済みのプロファイル、暗号化されたアイデンティティ、責任を負う所有者、明確な構造的関係をここで登録する必要があります。共有の人的認証情報を使用してチームが登録を回避することがないように、登録は十分に軽量なものにしてください。親のスコープを継承するローカルエージェントとエフェメラルなサブエージェントは、個々のディレクトリエントリではなく、以下に説明するコンテインメント制御と委任制御によって管理されます。エージェントがインポートされると、ディレクトリは、そのエージェントがエンタープライズでどのように動作するかについて、さらに2つの問いに答えます。それは、エージェントがどのような役割を果たすか、そして、どのように自身のアイデンティティを証明するかです。

エージェントクラス：どのような役割を果たすのか？

- **自律型エージェント**：人間の直接的な監視なしで非同期的に動作し、長期的なビジネス目標を達成する、完全に独立したエージェントです。
- **OBO (On-Behalf-Of) ヒューマンエージェント**：特定の人間ユーザーのセッションに直接紐付いたタスクを実行するエージェントです。代理として機能し、そのユーザーが持つ基盤の権限によって制約されます。
- **オーケストレーターエージェント**：目的をサブタスクに分解し、ダウンストリームのサブエージェントを生成、調整、監督する親レベルのエージェントです。ディレクトリは、委任チェーンと生成関係をファーストクラスの属性として追跡する必要があります。
- **サブエージェント**：親オーケストレーターによって動的に生成または呼び出される、特殊な子エージェントです。個別のディレクトリ登録ではなく、継承されたセッションスコープによって管理され、個別のサブタスクを実行します。LLM内で適用されるモデル内のタスク境界は失敗する可能性があるため、サブエージェントは、継承された権限スコープを構造的に適用する、分離された実行境界でラップする必要があります。これにより、推論エラーによる特権昇格を防ぎます。

- **物理/身体性エージェント**：ロボット、アクチュエーター、その他の物理システムを制御するエージェントです。このアーキテクチャはアイデンティティとアクセスを統制しますが、既存のプラント安全性、OTセキュリティ、変更管理の制御を置き換えるのではなく、補完するものです。
- **ガーディアンおよびセキュリティエージェント**：セキュリティコントロールプレーン内で稼働する、特化型の管理AIエージェント（SOCトリアージエージェント、動的ポリシーエンジン、継続的なAIASPMスキャナーなど）。テレメトリを検査し、封じ込めアクションを実行するための高い特権を持つため、ガーディアンエージェントのアイデンティティには、必須の暗号化ワークロード証明、厳格な職務の分離、および所有者による継続的なサインオフが必要です。

アイデンティティ基盤：それが本当にそのエージェントであることを何が証明するのか？

- **エージェントアイデンティティのメタデータプロファイル**：エージェントの意図、オリジン、リネージ、運用コンテキストを定義する豊富なメタデータプロファイル（CIMD）で、リスクベースのアクセス判断に情報を提供します。
- **ワークロードアイデンティティ**：動的なクラウド環境全体でエージェントに安全で検証可能なIDを発行するために使用される、暗号化され、プラットフォームに依存せず、標準ベースのアイデンティティ生成フレームワークです。

- **エージェントのガバナンスメタデータ**：エージェントのディレクトリレコードに紐付けられた構造化メタデータで、説明責任を負う人間の所有者、指定された運用チーム、またはオンコールキュー、ライフサイクルの状態（アクティブ、一時停止、廃止など）、リスク階層、運用スコープなどの情報が含まれます。このメタデータは、アクセス制御のオーサリングとランタイム評価のために、ポリシーエンジンによって継続的に使用されます。
- **A2Aエージェントプロトコル**：標準化された相互運用性フレームワークを活用して、動的に機能を検出し、タスクを委任し、エンタープライズの境界を越えてピアツーピアで連携するAIエージェントです。共有ディレクトリや信頼のルートがない組織間でやり取りする場合、エージェントは暗号による構成証明とトークン交換標準（WIMSE/AIMS、OAuth Token Exchange、HTTP Message Signaturesなど）を利用して、委任チェーンを維持し、仲介者間でプリンシパルのアイデンティティを検証します。

3.2 第2の柱：

エージェントは何ができるのか？ (アクセスとエンタイトルメント)

自律型エージェントは、本質的に行動するために設計されており、そのためにはデータ、API、システムに接続する必要があります。第1の柱で確立された検証済みアイデンティティを基盤として、この柱では、人間が直接委任したセッションと、複雑なマルチエージェントのワークフローの両方において、エージェントが企業リソースとやり取りする方法を統制する境界と権限を定義します。これにより、人間中心のレガシーモデルを、マシンツーマシンのAIガバナンスで強化します。

3.2.1 アクセスポリシー（アクセスポリシーの定義）

アクセスポリシーは、アイデンティティが検証されたエージェントが何を読み取り、書き込み、実行できるかを定める運用上の境界を定めたものです。生成AIの出力は予測不可能な性質を持つため、こうしたポリシーは、コンテキストに深く対応した、アダプティブなものである必要があります。

- **大まかな制御：**エージェントが接続できる高レベルの環境またはマクロサービスを定義する、ロールベースの広範な制御です。
- **きめ細かな制御：**属性ベースおよびオブジェクトレベルの権限により、エージェントを特定のデータベース行、ファイル、または特定のAPIメソッドに制限します。企業のデータ分類ラベル（機密、PIIなど）を明示的に考慮して、コンプライアンスに準拠したデータ境界を適用します。ポリシーの構成要素には、データ分類、エージェントのアイデンティティ、リスクのスコープ、環境コンテキスト、アクションの種類が含まれる場合があります。

- **関係性ベース：**エージェントのグラフ関係（生成元、誰の代理か、どの階層の下に属しているか）から派生するアクセス。
- **インテントベース：**リソースへのアクセスを付与する前に、エージェントのプロンプトやプランが意味する意図を分析する高度なポリシーエンジンです。現在のツールでプランを検査することは可能ですが、ユーザーのプロンプトをそのまま唯一のトリガーとして扱うのは理想論にすぎません。なぜなら、エージェントは必ずしもプロンプトだけに基いて行動するとは限らないからです。一時的かつ期限付きの特権は、特定のタスクの実行中のみを対象として作成され、タスク完了後すぐに取り消されます。
- **ネットワークベース：**エンタープライズVPC間のエージェントトラフィックを制限するマクロセグメンテーションと、エージェントとターゲット間の通信を承認済みIP範囲またはサービスメッシュに制限するL3/L4マイクロセグメンテーション制御を組み合わせたものです。

- **時間制限付き**：時間制限付きの権限付与は、単一のタスクまたはセッションにスコープが限定され、完了時またはタイムアウト時に自動的に取り消されます。自動承認、自動失効する権限付与だけでは、制御値は限定的です。そのため、有効期限切れのみを安全策として扱うのではなく、時間制限付きのスコープをセクション3.3.1.2で説明する異常検知と組み合わせて使用します。
- **アプリ間のアクセス**：エージェントが個別のアプリケーションエコシステム間を横断する能力を統制するポリシーフレームワークと境界制御です。これにより、エージェントが様々なエンタープライズプラットフォーム間でワークフローをオーケストレーションする際に、安全なデータ変換、トークン交換、および関数実行が可能になります。
- **委任と継承**：サブエージェントの効果的な権限境界は、混乱した代理問題のエスカレーションを防ぐため、サブエージェント自身のエンタイトルメントと、委任元エージェントまたは人間のエンタイトルメントの交差部分として計算されます。また、ポリシーでは、委任の幅と深度に上限を設けて、無制限な拡散を防ぐ必要があります。
- **ガードレール**：エージェントの意図やきめ細かい特権に関係なく、コンプライアンスに違反するアクション、有害な出力、リスクの高い運用手順をグローバルに防止するために、自律的な推論ループをオーバーライドする、厳格で決定論的な安全上の制約および行動上の境界です。

3.2.2 ガバナンス

ガバナンスは、コンプライアンスのベースラインと継続的な監視ループを提供します。手作業による形式的なレビューではなく、自動化されたツールを使用して、エージェントの特権が、エージェントの速度に合わせた最小権限の原則に沿ったものであるよう維持します。

- **条件付きレビュー**：エージェントのアクションが、付与された技術的な権限の範囲内であるかどうかにかかわらず、定義された財務上または運用上のしきい値（\$ 1,000を超える購入や予約など）を超えた場合に、自動または人間が介在するワークフロー承認をトリガーします。
- **アクセスレビュー**：自動的された継続的なエンタイトルメントのスキャンです。システム所有者が不適合な権限を検証できるよう支援します。
- **アクセスリクエスト**：既存のエージェントに新しい機能、モデル、またはシステムアクセスをプロビジョニングするための、正式で監査済みのワークフローです。リクエストは、人間の所有者から、または実行時にエージェントから動的に発行される場合がありますが、エージェントが開始したリクエストは、エージェントが自動化されたワークフローを操作して過剰なアクセス権を付与することを防ぐために、静的な権限の上限に照らして厳密に評価され、スコープを拡大する際には所有者による検証が必要となります。
- **職務分掌 (SoD)**：単一のエージェントが、金融取引の作成と承認の両方など、相反する特権を保有することを防ぐアルゴリズムベースのルールです。
- **入社・異動・退社 (JML)**：人間のアイデンティティライフサイクルで用いられるオンボーディング、変更管理、デプロビジョニングの規律を、エージェントのスコープ、所有権、またはモデルバージョンの進化にも同様に適用します。

3.3 第3の柱：

エージェントは何をしているのか？ (ランタイムの認可と制御)

インラインのランタイム監視と制御実施がなければ、可視性とアイデンティティは効果を発揮しません。この柱では、エージェントがタスクを実行する際にそのアクションをリアルタイムで傍受、検査、認可することに重点を置いています。これにより、エージェントが悪意のある行動を取ったり、機密データの取扱いを間違ったり、重要な企業資産を危険にさらしたりすることを防ぎます。

3.3.1 ランタイムの認可と監視

ランタイムの認可と監視は、ライブエージェントセッションのアクティブな検査レイヤーとして機能します。実行パスに直接配置され、エージェントへの入力と、エージェントから外へ拡散するアクションや出力の双方を分析します。

3.3.1.1 ゲートウェイ (アクセスポリシーを適用)

ゲートウェイは、多層防御チェーンにおいてポリシー適用ポイント (PEP) として機能します。エージェント型レイヤーとエンタープライズスタックの他の部分との間のトラフィックを傍受し、プロトコルを解析して、不正な呼び出しが宛先に到達する前にブロックします。ダウンストリームAPIゲートウェイは、アップストリームのAI/エージェントゲートウェイがすでに何を検査したかに関わらず、すべての呼び出しに対して独自の検証を実行し、最終的な認可権限を有します。AI/LLM、MCP、セキュリティ、エージェント、APIゲートウェイは、多くの場合まとめて展開されますが、それぞれが個別のコントロールポイントを所有しています：

AI/LLMゲートウェイはモデルAPIトラフィックを管理し、MCPゲートウェイはツール/コンテキスト共有プロトコルのトラフィックを管理します。また、セキュリティゲートウェイはプロトコルに依存しない脅威検査を提供し、エージェントゲートウェイはエージェント固有のアイデンティティと委任に関する課題を管理します。そして、APIゲートウェイは、エージェントが呼び出すツールやサービスに対して、従来のAPIレイヤーの制御を適用します。

- **AI/LLMゲートウェイ**：基盤モデルへの低レベルのAPI呼び出しを傍受し、レート制限、トークン数、システムプロンプトの上書きを管理します。
- **MCP (Model Context Protocol) ゲートウェイ**：MCPサーバーに重点を置き、エージェントとアプリケーション間のコンテキスト共有プロトコルとオープン標準の統合を管理するために最適化された、特化型プロキシです。
- **セキュリティゲートウェイ**：エージェントのやり取りに対して、専用のコンテンツセキュリティとプロトコル検証を提供するインラインの脅威軽減レイヤーです。トラフィックフローに直接配置され、TLSを終端してリアルタイムでペイロードを深く検査し、複雑なプロンプトインジェクション、ジェイルブレイク、隠された敵対的なテキスト文字列、悪意のあるコードペイロードがエージェントやダウンストリームのリソースに到達する前に積極的に傍受して取り除きます。また、データ分類スキャンを実行して外部に送るデータを編集したり、モデル出力の構造的整合性チェックや暗号化されたエージェントペイロードの暗号的検査を実行したりします。

- **エージェントゲートウェイ:** エージェント固有のランタイムコントロールプレーンとして機能します。すべてのリクエストにエージェントのアイデンティティを紐付け、エージェントが利用するツール/MCPレジストリを維持し、アイデンティティ、コンテキスト、およびリクエストされた特定のアクションに基づいて、アクセスだけでなくエージェントのアクションを管理します。安全でないインタラク션을リアルタイムで停止して、組み込みのポリシーとカスタムポリシーを介して、リスクのあるモデルトラフィックをブロックまたは無害化し、機密性の高いツールのアクションを拒否するか、または承認を要求します。また、実行時にシークレットレスで有効期間の短いダウンストリームの認証情報をAPI呼び出しに挿入することで、エージェントが永続的なエンタープライズAPIキーやサービスアカウントのシークレットを処理、保存、公開することがないようにします。サポート対象のマルチエージェントインタラクション全体で管理されたテレメトリを提供し、構成されたポリシーと機能に従ってPIIなどの機密データを検出または編集すると同時に、安全でない動作や悪意のあるプロンプトベースの攻撃から防御できます。
- **APIゲートウェイ:** エージェントとMCPサーバーが使用するツールやサービスに対して、従来のAPIレイヤーの制御、レート制限、クォータ、スキーマ検証、mTLS、WAF、トラフィックシェーピング、ゲートウェイでの取り消しを適用して、アップストリームのAI/エージェントゲートウェイがすべての異常をすでに検出したと想定するのではなく、すべての呼び出しに対して独自の最終認可チェックを実行します。

3.3.1.2 監視（ランタイムの監視）

監視機能は、詳細な可観測性、脅威検知、および能動的な制御適用を提供します。ランタイムのテレメトリをスキャンして、異常な動作、データ漏洩、外部からの脆弱性悪用の試みを検知し、サンドボックス化やエンドポイントの隔離といった制御をリアルタイムで適用します。

- **サンドボックス化:** エージェント（特に信頼できないエージェント、サードパーティエージェント、または新しく更新されたエージェント）の実行を、安全で制約のあるランタイム内に隔離することで、ラテラルムーブメントを防止し、構造的な境界によって実行リスクを封じ込めます。
- **エンドポイントの隔離と許可リスト登録:** OSレベルのサンドボックス化、エンドポイントアプリケーションの許可リスト登録、ローカルデバイスの態勢チェックなど、エンドポイントでホストされるエージェントのローカルランタイム制御により、ローカルエージェントのコマンド実行とファイルアクセスを制限します。
- **ツール/データアクセス:** ダウンストリームのツールやレジストリに渡される引数、ペイロード、スキーマ、クエリをリアルタイムで構造的に検査、監査します。実行境界の内部またはその直近で検査を行うことで、動的な実行プランが認可されたパラメーターから逸脱したり、セッションの途中で意図しない変更が加えられたりすることを防ぎます。
- **脅威/異常検知:** 機械学習エンジンがエージェントの行動ベースラインを追跡し、データアクセスにおける異常な急増、急速なAPIループ、または通常とは異なる実行パターンにフラグを立てます。

- **プロンプトインジェクション**: エージェントのコアロジックを乗っ取ったり、システム命令をオーバーライドしたりすることを目的とした、直接的および間接的な敵対的テキスト文字列を検出するように設計されたインラインスキャナーです。
- **DLP/出力フィルタリング**: データ損失防止 (DLP) ツールがエージェントの出力をスキャンし、独自のソースコード、個人を特定できる情報 (PII)、または企業秘密が外部プロバイダーや不正なユーザーに漏洩するリスクを軽減します。
- **トレース**: プライマリーユーザープロンプトによってトリガーされる正確な実行パス、コールスタック、サブエージェントの呼び出しを記録する、包括的なトランザクションマッピングです。
- **コストとパフォーマンス**: トークンの使用状況、運用コスト、レイテンシー、リソース消費量を追跡し、エージェントの無限ループやリソース枯渇攻撃を防ぎます。
- **ワークフローインザループ**: リスクの高いアクションに対してエージェントの実行を一時停止するプログラムによるトリガーです。実行を続行する前に、自動化されたポリシーのクロスチェックや明示的な人間による承認にルーティングします。
- **継続的な監視**: きめ細かいトレース、監査証跡のロギング、エージェントのアクションの否認防止など、エージェントのアクティビティを継続的に監視し、エクスポートする機能です。

3.3.1.3 継続的な意図と整合性の適用

継続的な意図と整合性の適用により、エージェントが実行中、委任されたタスクの範囲内にとどまっているか、または不正な目的拡大に向けてドリフトしていないかを判断します。これは初期アクセスポリシーを補完するものです。インテントベースのポリシーはエージェントが最初に何にアクセスできるかを決定しますが、継続的な整合性の適用では、変化するマルチステップの行動が長期にわたって適切であり続けるかどうかを評価します。

主な機能は次のとおりです。

- **目的のバインディング**: 委任元のユーザー、エージェントのアイデンティティ、承認済みのワークフロー、要求されたタスクの間で厳密な連携を徹底します。
- **計画とアクションの評価**: 計画されたツール呼び出しと観測されたアクションをランタイムで比較し、タスクのドリフト、インジェクションによる逸脱、および混乱した代理人問題を検出します。
- **コンテキストに対応する適用**: 許可、拒否、修正、ステップアップ承認、スコープの縮小、一時停止、封じ込めなどのアクションをリアルタイムで実行します。
- **追跡可能な意思決定の証拠**: 重要なアクションが許可、変更、またはブロックされた理由を説明する、否認防止機能を備えた監査ログを維持します。

3.3.2 リソースアクセス

リソースアクセスは、エージェントの実際の影響範囲を明確に示します。エージェントがやり取りできる多様なダウンストリームターゲットを分類し、すべてのインターフェイスが適切なセキュリティラッパーで保護されるようにします。

3.3.2.1 ターゲットリソース（アクセスするコンポーネント）

ターゲットリソースは、エージェントのアクションの宛先を表します。コンテキストのプルと変更のプッシュのどちらを実行する場合でも、こうしたエンドポイントを継続的にカタログ化して防御する必要があります。

- **MCPサーバー**：Model Context Protocol (MCP) を介してエージェントとコンテキスト情報（構造化データ、ドキュメント、埋め込み、ツール、状態）を共有するように明示的に構成されたデータリポジトリおよびユーティリティサーバーです。
- **SaaSアプリ**：機密性の高い企業データを大量に保存する、エンタープライズの生産性向上および基幹業務アプリケーションです。
- **データストアとナレッジリポジトリ**：ベクトルデータベース、データレイク、ウェアハウス、トランザクションストレージシステムなど、構造化および非構造化リポジトリです。エージェントはこれらを利用してビジネスコンテキストを抽出したり、運用上の出力をコミットしたりします。これらのリポジトリは、データセキュリティ態勢管理 (DSPM) プラクティスによって管理されており、エージェントが各ストアにアクセスする際にリネージと分類を追跡します。

- **CLIツール**：エージェントがシステムレベルのスクリプト、構成、または運用コマンドを実行できるようにするコマンドラインインターフェイスです。
- **その他のエージェント**：より広範なワークロードの一部の構成要素を実行するよう割り当てられた、ダウンストリームのサブエージェントまたは特化型エージェントです。
- **特権付きリソース**：ルートディレクトリ、アイデンティティストア、マスター構成コンソールなど、機密性の高いターゲット環境です。
- **標準エンドポイントとレガシーエンドポイント**：MCPなどのエージェント型コンテキストプロトコルをネイティブにサポートしていないエンタープライズ REST/GraphQL API、データベース、レガシーアプリケーション。
- **決済システム**：金融ネットワーク、トランザクションゲートウェイ、会計台帳には、最高レベルの暗号検証と人間による監視が求められます。
- **ブラウザ**：エージェントが公開情報をスクレイピングしたり、Webベースの外部コンソールを操作したりするために利用する、Webの自動化されたセッションです。
- **スキルのランタイムと実行可能コード**：コードインタープリター、スクリプト実行サンドボックス、およびエージェントによって動的に呼び出されて生成コードを実行するプロシージャルモジュールで、ランタイムエスケープのリスクを軽減するために、隔離されたコンピューティング境界を必要とします。
- **基盤モデル**：重みがホストされる未加工のインテリジェンスプロバイダーで、安全で検証済みのモデルAPIキーを必要とします。
- **コミュニケーションおよびコラボレーションチャネル**：エージェントがやり取りを行うメール、チャット、チケットシステムで、ペイロードの深い検査とスキャンの対象となる、信頼できないプロンプトの入力パスを表します。

3.3.2.2 目的に応じたデータセキュリティ

目的に応じたデータセキュリティでは、委任された目的、ユーザーの権限、エージェントのアイデンティティ、ワークフローの状態、データの分類、受信者、宛先、ダウンストリームのアクションに基づいて、特定のデータを特定のエージェントがアクセス、処理、変換、要約、開示、または送信すべきかどうかを評価します。アクセス時だけでなく、使用時にもデータ制御を適用します。エージェントは、システム間でデータを組み合わせ、機密性のない入力から機密情報を推測し、出力、API、メッセージ、ファイル、サブエージェントの呼び出しを通じて結果を開示するため、この点は重要です。データの分類とDLPによって、そのデータが何であるか、機密情報であるかどうか判断されます。一方、目的に応じたデータセキュリティによって、その状況でそのデータを使用することが適切かどうか判断されます。主な機能は次のとおりです。ユーザーのリクエスト、エージェントのタスク、データソース、許可された用途の間の目的の紐付け。アクセスと開示が現在のワークフローのステップに適しているかどうかのランタイム評価。要約、変換、エクスポート、アプリ間の移動、サードパーティのエージェントへの開示に対する制御。データの過剰な取得、過剰な検索、不正な推論、不適切な共有の検知。編集から、マスキング、部分的な応答、承認ワークフロー、ログ記録、ブロック、封じ込めにわたる強制適用。

3.4 第4の柱：

どのように対応すべきか？ (アクティブな封じ込め)

このブループリントの最後の柱は、テレメトリーを即時の封じ込めに変えることです。監視システムが異常な動作、認証情報の侵害、またはプロンプトインジェクション攻撃を検出した場合に備えて、エンタープライズには、自社の対応によって停止を引き起こしかねない破壊的なアクションを取るのではなく、問題の規模に応じて影響を最小限に抑えた対応（レート制限や隔離など）を優先する、高度に自動化された精緻な緩和メカニズムが必要です。

3.4.1 対応/強制適用（スロットリングと封じ込め）

自動化された運用制御ループは、リスクシグナルにリアルタイムで対応するように設計されています。これは単一の一元的なチョークポイントとしてではなく、連携した制御レイヤーとして機能し、技術スタック全体にわたって、適切なセキュリティレスポンスを実行します。

強制適用は、標準化された通知プロトコルと組み合わせることで最も効果を発揮します。つまり、リスクを検出したシステムがキルスイッチイベントを発行し、サブスクライブしているシステムがそれに基づいて動作し、その結果を確認シグナルとして返すシステムです。

3.4.1.1 適用アクション（軽減メカニズム）

適用アクションとは、セキュリティチームや自動化されたオーケストレーターが、問題のあるエージェントや侵害されたエージェントを隔離、封じ込め、または完全に無力化するために実行できる、具体的かつ多層的な手段のことです。これらのアクションは、隔離された単発のスイッチではなく、SOCが所有する対応オプションのライブラリを形成します。対応担当者は、集約されたリスクシグナルに応じて、最も影響の少ないアクションを選択します。

ガーディアンエージェントによってトリガーされる封じ込めアクション（一時停止、分離、終了）では、ターゲットエージェントが、マネージドクラウドサービス、SaaSアプリケーション、コンテナ化されたランタイムのいずれで実行されているかに関わらず、きめ細かいセッションレベルの制御をサポートする実行境界内で動作する必要があります。隣接するワークフローに付随的な影響を与えることなくエージェントを一時停止するための、基盤となるプラットフォームやランタイムのサポートがなければ、API主導の封じ込めは誤った安心感を与えることになります。物理的な機器やOTに関連付けられたエージェントを封じ込める場合、突然の隔離によって工場の現場で物理的な安全上の危険や運用上の危険が生じないように、自動化されたトリガーを調整する必要があります。

- **トークンの失効：**アクティブな OAuth トークンまたは API アクセストークンを即座に無効にし、エージェントが接続先のアプリケーションを呼び出す機能を直ちに停止させます。
- **セッショントークンの失効：**エージェントの広範なアイデンティティを必ずしも削除することなく、アクティブな OAuth トークンを無効にし、接続されているエンタープライズサービス全体でダウンストリームのセッションコンテキストを失効させる、または単一の特定のインタラクション文字列のアクティブなランタイム接続を切断します。
- **継続的な認可：**リスク態勢が変化した瞬間に権限を即座にダウングレードする、セキュリティコンテキストの動的かつリアルタイムな再評価です。

- **プロセスの終了:** 隔離に失敗した場合の最終手段として、コンピューティングランタイムを強制終了します。ただし、終了するとメモリ内のフォレンジックテレメトリが失われるため、可能な限りネットワーク隔離、セッションの中断、または実行環境の中断を選択してください。
- **ネットワークの隔離:** マイクロセグメンテーションのルールを利用してエージェントのホスティング環境を隔離し、内部ネットワークとオープンなインターネット双方との通信を遮断します。
- **クラウドサービスの終了:** 不正なエージェントをホストしている基盤のクラウドインスタンス、サーバーレス関数、または仮想マシンクラスターを即座にシャットダウンします。
- **エージェントプラットフォームの終了:** エージェントの運用アカウントを一時停止し、スケジュールされたすべてのタスクを停止するよう、一元化されたエージェントオーケストレーションプラットフォームに指示する、トップレベルの管理コマンドです。
- **スロットリングとレート制限:** 適度な初動対応として、エージェントのリクエストレート、トークンバジェット、または同時実行性を段階的に制限し、セッションを完全に切断することなく機能を低下させます。

3.4.2 アクセス復旧

キルスイッチは制御ループの半分にすぎません。すべての封じ込めアクションには、検証済みで既知の直近の正常状態に戻すための、対応するパスが必要です。アクセス復旧は、封じ込め、隔離、または終了されたエージェントをサービスに復元する方法を規定し、強制措置が自動的に取り消されるのではなく、復帰が意図的で、文書化され、追跡可能であることを確認します。完全に自動化・標準化された復旧ワークフローは依然として目標の状態であり、今日のほとんどの組織では、以下のいずれか1つ、またはそれ以上のステップで手動による承認を必要としています。

- **再構成証明:** 新しく隔離された実行インスタンスでアクセスを復元する前に、既知の正常なベースラインと照合してエージェントのアイデンティティ、コード、構成を再検証します。
- **段階的な再登録:** すべての特権を一度に復元するのではなく、アクセス権を段階的に復元します（例：書き込み権限の前に読み取り専用権限、本番環境の前にサンドボックス環境）。
- **根本原因の承認:** エージェントを再度有効にする前に、担当者またはセキュリティチームに根本原因と修復策を文書化するよう要求します。
- **監査証跡のクローズ:** 元のキルスイッチイベント、修復アクション、再登録の決定を、クローズした監査可能な1つのインシデント記録に紐付けます。

3.5

横断的な基盤コンポーネント

これらのアーキテクチャレイヤーはサイロ化されているわけではなく、4つの柱すべてにまたがる横断的な基盤を提供し、可視性、アイデンティティ、監視、対応を1つのセキュリティファブリックにつなぐ結合組織として機能します。

3.5.1 実行コンテキストとリスクシグナル

実行コンテキストとリスクシグナルは、このブループリントの要素をつなぐ分析的な組織として機能します。リアルタイムのランタイムテレメトリ、セキュリティ態勢評価、および委任元のユーザーアイデンティティシグナル（アカウント侵害のインジケータや高権限の主体リスクなど）から継続的に情報を受け取るこのオーケストレーションエンジンは、個別では低リスクの状態でも相関的に有害な組み合わせとなる脅威を動的に評価します。問題が発生した場合のアラートだけでなく、何が有効かに関する既知の構造的な確認情報、つまり正常な実行コンテキストも公開します。しきい値を超えた場合（例えば、唐突なプロンプトインジェクションと財務データ窃取の試みの組み合わせなど）、リスクシグナルが対応する対応/強制適用メカニズムを直接的かつプログラマ的に作動させます。異種ベンダー環境全体でリアルタイムの相互運用性と協調的な対応を確保できるよう、このシステムは **Shared Signals Framework (SSF)** や **Continuous Access Evaluation Profile (CAEP)** などのオープンなセキュリティエンジニアリング標準を活用して、ゼロトラストの状態変更を非同期で公開および利用できます。

3.5.2 AIエージェントのテレメトリ、ログ、可観測性

これは、リファレンスアーキテクチャのすべての柱を支える、基礎となるベースラインレイヤーです。これは不変のデータプレーンとして機能し、エンタープライズスタック全体で実行メタデータをキャプチャします。また、必須のインラインペイロード編集によって、PII、シークレット、認証情報がログレイクに入るのを防ぎます。このような構造化された高精度のロギングレイクがなければ、自動化された脅威検知は不可能になり、態勢評価にはデータが不足し、フォレンジックインシデントレスポンスはまったく状況を把握できなくなります。複雑なマルチエージェントインタラクションにおける一貫性を確保し、ベンダーのサイロ化を軽減するために、このレイヤーでは **Open Cybersecurity Schema Framework (OCSF)** などのオープンな標準を活用して、多様なログストリームを単一の統一されたセキュリティ分類に正規化できます。

- **改ざん防止テレメトリのオリジン:** 自己報告型のエージェントログに依存するのではなく、実行境界でテレメトリをキャプチャすることで、否認防止を徹底できます。これにより、侵害されたエージェントが実行記録を隠蔽、変更、改ざんすることを防ぎます。

4

グローバルシステムインテグレーター（GSI）との連携による運用

アーキテクチャのブループリントには、永続的なエンタープライズ価値を実現するための運用における展開モデルが必要です。以下のセクションでは、セキュリティチームとアドバイザリーパートナーが、既存のエンタープライズワークフロー内でこれら4つの柱をどのように運用に落とし込むかを定義します。組織は、その能力や要件に応じて、社内のエンタープライズチーム、GSI、マネージドサービスプロバイダー、またはその他の認定パートナーを通じてこのアーキテクチャを運用できます。運用モデルなしで公開されたリファレンスアーキテクチャは、1年以内に使われなくなってしまう。以下のセクションでは、そのアーキテクチャをエンタープライズ内で実際に立ち上げ、人員を配置し、維持していく方法を定義します。

4.1 デプロイメントアドバイザリーとアーキテクチャイネーブルメント

GSIは、4つの柱を一斉に同時展開するのではなく、組織毎に順序立てられたロールアウトとして実行します。これには以下が含まれます。

- **現状評価:** エンタープライズの既存のIAM、ネットワーク、監視スタックを「Blueprint」のリファレンスアーキテクチャにマッピングし、すでに別の名称で存在する機能と比較して真のギャップを特定します。

- **ターゲットオペレーティングモデルの設計:** 各柱における運用上の明確な所有権を定義し、チーム間の責任の所在が不明確になることを防ぎます。このモデルでは通常、ゲートウェイと監視をセキュリティエンジニアリングに、エージェントディレクトリ、アクセスポリシー、ガバナンスをIAMに、対応/適用をSOCに、実行境界（サンドボックス化、隔離ランタイム、ライフサイクル管理）をコンピューティングおよびプラットフォームエンジニアリングに割り当てます。
- **段階的なロールアウト計画:** クライアントの既存のツールへの投資と規制上の期日に合わせて、導入の順序を決定します。
- **ベンダーの相互運用性の統合:** 企業がすでにライセンスを取得している特定のAI/LLM、MCP、セキュリティ、エージェント、APIゲートウェイを、既存のインフラストラクチャを全面的に刷新することなく、共有テレメトリおよびリスクシグナルレイヤーを介して相互運用できるように構成します。

4.2 標準化されたリスクスコアリングマトリクス

すべてのエージェントに同じレベルの多層制御が必要になるわけではありません。GSIは、組織が低リスクの展開に過剰な対策を講じたり、高リスクの展開で保護が不十分になったりするのを防ぐ、適切に調整されたスコアリングフレームワークを提供します。これには以下が含まれます：

- **リスク階層化モデル：**エージェントを運用スコープ（読み取り専用かトランザクションか）、データの分類、自律性レベル（OBOか完全自律型か）、ターゲットリソースの影響範囲に照らしてスコアリングします。
- **業界に合わせて調整されたしきい値：**決済ワークフローを実行する金融サービスのエージェントと、OTセンサーデータを読み取る製造業のエージェントとでは、リスクのベースラインが大きく異なります。GSIは、業界横断的なパターンライブラリを活用して、こうしたしきい値をゼロから設定するのではなく、適切に調整します。
- **セマンティックガバナンスのためのビジネス影響スコアリング：**汎用的なデフォルト設定ではなく、クライアントの実際のリスク許容度に合わせて、人間の介在による承認を必須とする金額、規制、または評判に関するしきい値を定義します。
- **リスク階層に応じた分離レベル：**リスク階層に応じて、レビューによるゲート設定と並行して、実行環境を構造的に隔離する仕組み（カーネルレベルのサンドボックスや一時的なランタイムライフサイクルなど）の強度も調整する必要があります。これにより、高リスクのエージェントが侵害された場合に、低リスクのツールと同じ物理的な範囲に影響が及ばないようにすることができます。

4.3 継続的なライフサイクル管理

エージェントのガバナンスは、一度きりの認定ではありません。GSIは、環境の変化に応じてエージェントディレクトリとアクセスポリシーを正確な状態に維持するための継続的なライフサイクルプロセスを運用します：

- **アクセス認定サイクルの自動化：**マネージドサービスがエージェントディレクトリとアクセスポリシーの状態を取得し、担当である人間の所有者が組織を離れたエージェントや、最小権限から逸脱したエンタイトルメントを持つエージェントにフラグを立てます。
- **エージェントの入社・異動・退社（JML）プロセスの同等運用：**人間のオフボーディングで用いられるデプロビジョニングの規律を、エージェントの所有権変更、モデルバージョンのアップグレード、サブエージェントの廃止にも同様に適用します。これにより、孤立した認証情報が蓄積してアイデンティティギャップの主な原因とならないようにします。
- **職務分掌（SoD）例外ルーティング：**人間のアイデンティティガバナンスと同じ頻度で、競合する権限の例外（金融取引の作成と承認の両方が可能なエージェントなど）を、解決のために適切なシステム所有者にルーティングするワークフローを運用します。
- **実行環境のドリフト管理：**エージェントのランタイムと基盤となるプラットフォーム構成に対して、定期的なパッチ適用、依存関係の更新、再証明を実施して、ホスト型、クラウド、コンテナ型のデプロイメント全体で脆弱性のドリフトを解消します。

4.4 規制およびコンプライアンスとの整合

GSIは、企業が事業を展開する各地域の具体的な規制体制とアーキテクチャを橋渡しして、技術的な制御を、規制、コンプライアンス、監査の要件に対応するために設計された証拠へと変換します。

- **規制フレームワークへの制御マッピング:** エージェント型AIに特化したガイダンスの進化に合わせて、テレメトリ、監査証跡、アクセスレビューの出力を、業界固有の要件（金融サービスのモデルリスク管理、ヘルスケアのデータ処理、重要インフラの保護など）に対応付けます。
- **監査に対応した証拠パッケージ:** テレメトリおよびロギングレイヤーの出力を構造化し、アクセスレビュー、封じ込めとスロットリングのイベント、アクセス復旧の記録（再証明、根本原因の承認、監査証跡のクローズ）が、社内のSOCのニーズだけでなく、審査官や監査人の文書化基準も満たすようにします。
- **越境データ移転とデータレジデンシーに関するガイダンス:** エージェントのインポートチャネル（SaaSエージェント、組織の境界を越えるファーストパーティエージェント）が、エンタープライズが事業を展開する各管轄区域に固有のデータレジデンシーおよび越境データ移転の要件とどのように関連するかについて助言します。
- **改ざん検知が可能な監査来歴:** テレメトリ収集を、規制当局や検査官が実行境界で直接生成された否認不可能な監査ログで受け取れるように構造化し、周辺のネットワークインフラから二次的なログを再構成する必要性を減らします。

4.5 業種別のプレイブック

制御要件は業種によって異なるため、GSIはそれぞれ異なる運用環境に合わせたプレイブックを維持する必要があります。

- **金融サービス:** 厳格な実行分離、保証レベルの高いトランザクションしきい値、規制データの制御が重視されます。
- **製造・工業:** レガシーハードウェア全体におけるOT、IoT、エッジエージェントの検出が重視されます。
- **小売・コマース:** B2Cのプロンプトインジェクション対策、ブランド保護、顧客のPII漏洩に重点を置いています。

創設メンバーは、共同の相互運用性検証の結果が公開されるたびにこれらのプレイブックを継続的に更新し、新たなエージェント型機能の登場に合わせて運用モデルも適応していることを検証しています。

まとめ：AIセキュリティの未来の形成

Blueprint Allianceは、本番環境のランタイムにおけるエージェントの動作を統制します。外部レイヤーとして後付けされるのではなく、エンタープライズスタックに直接組み込まれたこのアーキテクチャは、エージェント型自動化を安全に大規模導入するために必要な構造的安全対策を提供します。

真のマルチベンダー相互運用性を実現するため、創設メンバーは、ツール間のインタラクションのための Model Context Protocol (MCP)、ログ統合のための Open Cybersecurity Schema Framework (OCSF)、リアルタイムのリスク情報交換のための Shared Signals and Events (SSF/CAEP)、リクエスト認証のための HTTP Message Signatures (RFC 9421) など、中核となるオープン標準に照らして自社のプラットフォームを積極的に検証しています。この共有テレメトリベースラインにより、あるランタイム環境で検出された脅威シグナルや境界違反を、コントロールプレーン全体でネイティブに取り込み、対応できるようになります。機能はベンダーや実装によって異なります。リファレンス統合と検証済みの相互運用性の結果は、テスト完了次第公開されます。

Allianceは、共同の相互運用性検証の結果やリファレンス統合を定期的に公開することで、エージェント型イノベーションの速度に合わせて進化する、マルチベンダーのエコシステムを実現します。

Blueprintの実践

以下の5つのシナリオは、セクション3で説明した各柱が実際にどのように連携して機能するかを説明するものです。いずれも、特定の顧客1社を対象としたものではなく、Allianceの初期の取り組みで見られたパターンを組み合わせた複合的なシナリオです。これらの複合事例は、このアーキテクチャが想定される動作を示すものであり、現時点でAllianceのメンバー規模で実際に導入された成果を示すものではありません。1つの導入事例ではなく、アーキテクチャが実際に動作する様子を示すものとして捉えていただければと思います。

シャドーエージェントから登録済みアイデンティティへ。財務アナリストが、ベンダーとの契約を要約するために、ブラウザベースのAIアシスタントを自社のSSOセッションに接続したとします。シャドーAIディスカバリーは、未認識の外部トラフィックに対して、最初の接続時にフラグを立てます。AIエージェントのセキュリティ態勢管理は、そのツールに既知の脆弱性がないかを確認し、エージェントディレクトリにルーティングします。機密データへのアクセスが行われる前に、そのアシスタントは、検証済みのメタデータプロファイルと範囲が限定されたアクセス権を持つ管理対象アイデンティティとして登録されるか、ネットワークレイヤーでブロックされます。これまで何か月も見えないままだったものが、機密データに触れる前に可視化され、確認され、対応が決定されるようになります。

マルチエージェントワークフロー全体で委任を維持。セールスオペレーションの責任者が、一連のCRMレコードを更新し、配布リストに通知するようオーケストレーターエージェントに依頼したとします。オーケストレーターは、書き込みを処理するサブエージェントと、通知の下書きを作成する別のサブエージェントを生成します。各ホップでは、ネストされたアクタークレームが元の人間のセッションにまで遡って引き継がれるため、エージェントディレクトリとアクセスポリシーエンジンは、各ステップでOBOの制約を適用できます。サブエージェントが、元の間がアクセスできないシステムにアクセスしようとする、Cross App AccessなどのコアIAMとポリシーの境界によってブロックされます。また、各サブエージェントは、独自の分離された実行コンテキストで動作するため、ブロックされた攻撃がオーケストレーターのセッションや兄弟関係にあるタスクに影響を与えることはありません。一方で、ワークフロー全体は単一の実行トレースに統合された状態で維持されます。

エージェントが実際に行おうとしている操作に合わせて、アクセス権を適正化。サポートエージェントは、顧客データプラットフォームへの広範な読み取りアクセス権でプロビジョニングされています。このエージェントが返金を処理しようとする、インテントベースのポリシーエンジンが、照会から取引処理への移行を検知し、その1回の注文に限定されたきめ細かいエフェメラルな特権を付与する前に、人間の介入による承認を要求します。承認されたその1件のタスクにスコープを限定したエフェメラルなランタイム内で返金処理を実行することで、タスク完了時に昇格された特権とランタイム環境が確実に廃止されるため、永続的な権限と持続的な攻撃対象領域が削減されます。昇格された特権は、タスクが完了した瞬間に失効します。

異常の検知から封じ込めの自動化まで。ランタイムの監視（通常はLLMではなく従来のML異常検知モデルによって行われる）によって、エージェントがほとんどアクセスしない社内APIに異常に大量の呼び出しを行っていることが検知されます。一方、インラインのGuardian AI Agentは、直近の入力に含まれているプロンプトインジェクションの兆候に、個別にフラグを立てます。Risk Signalsは、Shared Signals Frameworkを通じて両方のイベントを相関分析し、定義されたリスクしきい値を超えると、トークンの失効、ネットワークの隔離、エージェントのアクティブな実行環境の停止を自動的にトリガーします。さらに、フォレンジック調査に備えてメモリ内状態を保持し、実行環境を完全に終了させないようにします。これらすべてを、人がアラートに気付くのを待たずに実行します。セキュリティチームは、対応完了を確認できる完全なトレースを受け取り、封じ込めに要する時間を数時間におよぶ手動調査から自動化されたほぼ即時の対応へと短縮できます。

アクセスレビューを継続的なプロセスへと転換。継続的なアクセス認定サイクルごとに、グローバルシステムインテグレーターのマネージドサービスは、Blueprintのオープンなテレメトリレイヤーを介してエージェントディレクトリとアクセスポリシーの状態を完全に把握し、所有者が退職したエージェントにフラグを立て、職務分掌の例外を適切なシステム所有者にルーティングします。スプレッドシートを使って手作業で行っていたレビューが、監査に対応した継続的なワークフローに変わります。

利用開始に向けたベストプラクティス

このBlueprintに含まれるすべての機能を初日から利用する必要はありません。Allianceメンバーが実際に導入した際の順序は次のとおりです。

最初に、統合ログとテレメトリーを確立します。単一の正規化されたログレイクがなければ、ポリシーの適用やベースラインの動作をデバッグすることはできません。これを、以下のすべてのステップの前提条件として扱ってください。

次に、ポリシーではなく、ディスカバリーから始めます。見えないものは統制できません。アクセスポリシーを1つ作成する前に、シャドー AIのディスカバリーと初期のエージェントインベントリーを確立します。ほとんどのチームは、すでに稼働しているエージェントの多さに驚きます。

制限する前に、登録します。きめ細かなポリシーやインテントベースのポリシーを追加する前に、検出したすべてのエージェントを、検証済みのアイデンティティや所有者情報と一緒に単一のエージェントディレクトリに登録します。未登録のエージェントに対するポリシーは、適用することすらできません。

認可を階層化します。インテントベースの認可によってエージェントの権限を適切なスコープまで絞り込むには、その前提として、きめ細かな認可 (FGA) と関係ベースのアクセス制御 (ReBAC) が備わっている必要があります。

対応を自動化する前に、ランタイムを計測します。リスクシグナルに自動化された適用措置を連携させる前に、監視、トレーシング、DLP/出力フィルタリングを広範囲に導入して、実際の動作ベースラインを確立します。ノイズが多い、または不完全なテレメトリに基づいて封じ込めを自動化すると、誤検知によってシステム全体の信頼性が損なわれます。

キルスイッチは、必要になる前に負荷テストを実施してください。トークンの失効、セッション終了、ネットワーク隔離を、非本番環境のエージェントに対して定期的に検証します。テストされていないキルスイッチは仮説にすぎず、制御手段とは言えません。

ガバナンスはプロジェクトではなく、継続的なプロセスとして扱います。アクセスレビュー、アクセスリクエスト、職務分掌のチェックは、エージェント型AIに関する1回限りの取り組みとしてではなく、人間のアイデンティティガバナンスと同じ頻度で実行する必要があります。グローバルシステムインテグレーターは、プロジェクトから運用モデルへのこの移行を支援します。

これらの資料は、一般的な情報提供のみを目的としており、法律、プライバシー、セキュリティ、コンプライアンス、またはビジネス上の助言ではありません。本コンテンツは、セキュリティ、法律、規制に関する最新の動向を反映していない可能性があります。このコンテンツの利用者は、自身の責任において、自身の法のおよび/または専門的アドバイザーから助言を得るものとし、本資料を専門家による助言の代わりとして依拠すべきではありません。著者は、本コンテンツに関していかなる表明または保証も行いません。また、これらの推奨事項をお客様が実装した結果として生じたいかなる損失または損害についても責任を負いません。