AIガバナンスの3層管理モデル

情報の機密性とリスク許容度に基づく実践的導入ガイド

NIST AI RMF / 経済産業省ガイドライン準拠・技術的/運用的境界によるセキュアなAIアーキテクチャの構築

The Core Challenge:
イノベーションの加速とシャドーAIの蔓延

The Drive

加速するビジネス要求

- 生産性向上の強い圧力
- 現場主導でのAIツールの急速な普及

The Risk

見えないリスクの増大 (Shadow AI)

- **Stage 2: Workflow Embedding**: 承認なしに日常業務（コード記述、ドキュメント要約）にAIが組み込まれる「ガバナンスの空白地帯」
- **Stage 3: Risk Blindness**: 機密コードや顧客データのパブリックモデルへの無自覚な入力。データ漏洩とコンプライアンス違反の時限爆弾。

「ゼロトラストの時代において、
AIエージェントの自律的アクション (API呼び出し、
データアクセス) をどう制御するか？」

The Governance Foundation: 抽象的理念から具体的管理策への変換

理念・原則 (Principles)

AI事業者ガイドライン（総務省・経産省）：
人間中心、プライバシー保護、セキュリティ確保など10の共通指針。

プロセス (Processes)

NIST AI RMF（100-1 / 600-1）：AI特有のリスク（作話、有害コンテンツ、プロンプトインジェクション等）を管理する4機能（MAP, MEASURE, MANAGE, GOVERN）。

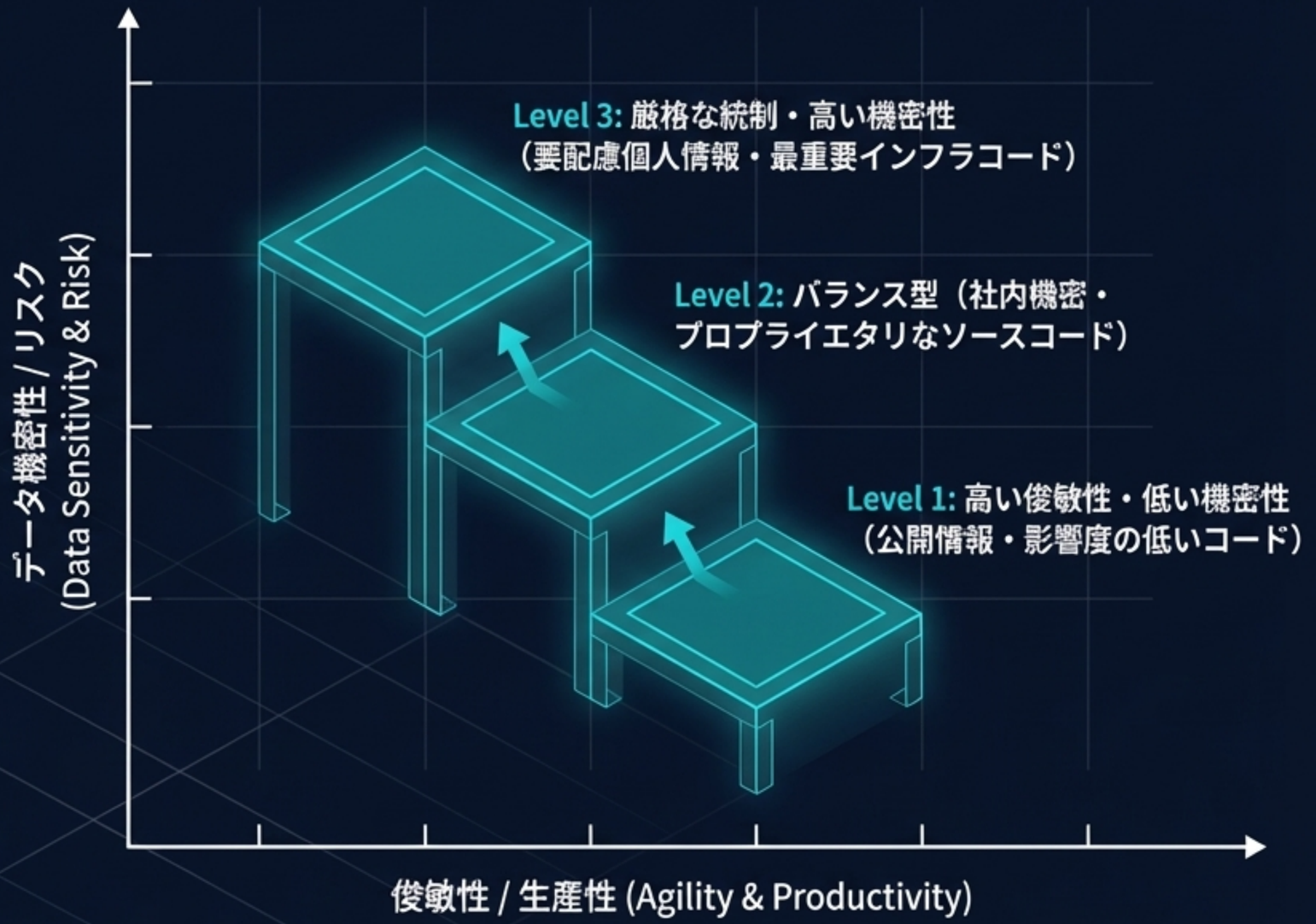
管理策 (Controls)

OWASP Top 10 for LLMs / MITRE SAFE-AI / CSA AICM：脅威のモデリング（40種類）と、アクセス制御・サンドボックス化などの具体的な技術的緩和策。

Key Insight:「ガイドラインを読むだけではシステムは守れない。理念をアーキテクチャに落とし込む『設計図』が必要である。」

The Core Thesis: 機密性とリスクに応じたアーキテクチャの分離

AIセキュリティは「ゼロかヒャクか」ではない。
過度な制限はイノベーションを殺し、野放しは組織を破壊する。
扱うデータの機密レベルに応じてシステム設計を動的に変化させる「3層管理モデル」が最適解となる。



Architectural Blueprint: AIガバナンスの「3層管理モデル」

Layer 3: Regulated / Enterprise (Level 3)
- 対象: 金融・医療などの規制業種、大企業
- フォーカス: 厳格なデータ主権。完全閉域網とゼロデータ保持（ZDR）による最高レベルの堅牢性。

Layer 2: Mid-Size / Standard Enterprise (Level 2)
- 対象: 中規模企業、標準的な開発組織
- フォーカス: 組織的統制。ID管理（SSO）とOSレベルの隔離による包括的制御。

Layer 1: Startup / Agile Team (Level 1)
- 対象: スタートアップ、小規模アジャイルチーム
- フォーカス: スピードと俊敏性。ツール設定と運用ルールによる軽量な防御。

Level 1: スタートアップ / アジャイル環境

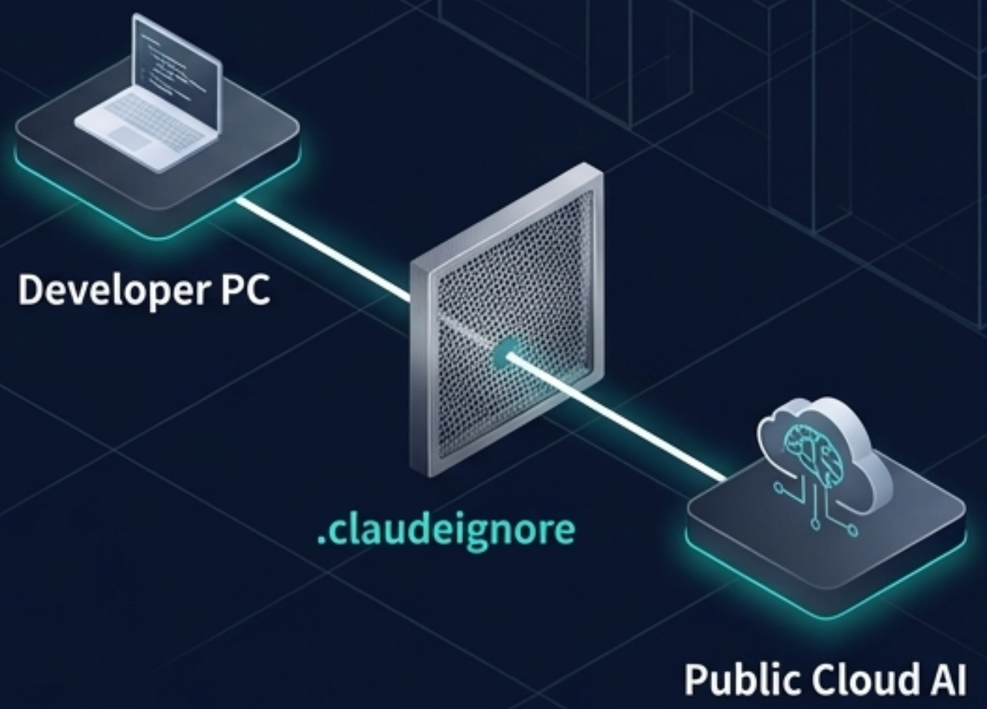
俊敏性を最優先し、ローカル制御とポリシーで防御する

推奨インフラ（Infrastructure Component）

- Direct API / 商用プラン利用：「商用利用で学習に使われない」契約（Anthropic商用利用規約等）を明示的に確認し利用。

防壁・境界（Governance & Boundaries）

- Technical Boundary (技術的境界) : .claudeignore ファイルの設定による機密ファイル（.env, *.pem, credentials等）のAIからの隠蔽。settings.json の denyルールによる危険なコマンド（curl, ssh等）の実行ブロック。
- Operational Boundary (運用的境界) : 最低限のポリシー策定と開発者へのAIリテラシー教育（プロンプトに顧客データを入れないガイドライン運用)。常にAIツールを最新版に保つ (脆弱性 CVE 対策)。



Level 2: 中規模組織 / 標準エンタープライズ

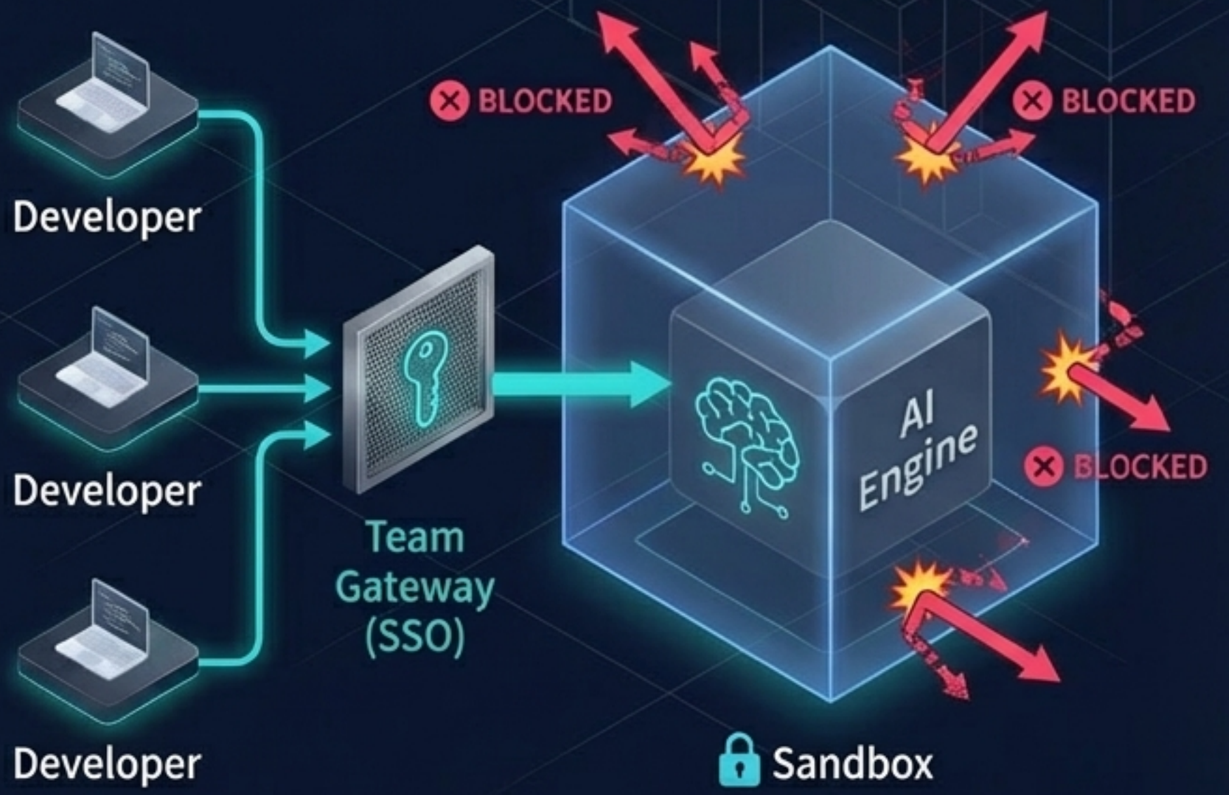
チームの交差を管理し、OSレベルの隔離と権限管理を導入する

推奨インフラ（Infrastructure Component）

- 組織向けプラン利用 : Claude Team、GitHub Copilot Businessなど、組織管理が可能なSaaS環境。

防壁・境界（Governance & Boundaries）

- Technical Boundary (技術的境界) : OSレベルのサンドボックス環境（macOSのSeatbelt、Linuxのbubblewrap等）の有効化。ファイルシステムとネットワークを強制隔離し、プロンプトインジェクションや攻撃者ドメインへの通信を防御。
- Operational Boundary (運用的境界) : SSO/SAML連携による一元的な権限管理。退職・異動時の即時アクセス停止（Need to Know原則の徹底）。AI資産のインベントリ管理。



Level 3: 規制業種 / 大規模エンタープライズ

完全なデータ主権と、監査に耐えうる厳格な証跡管理

推奨インフラ (Infrastructure Component)

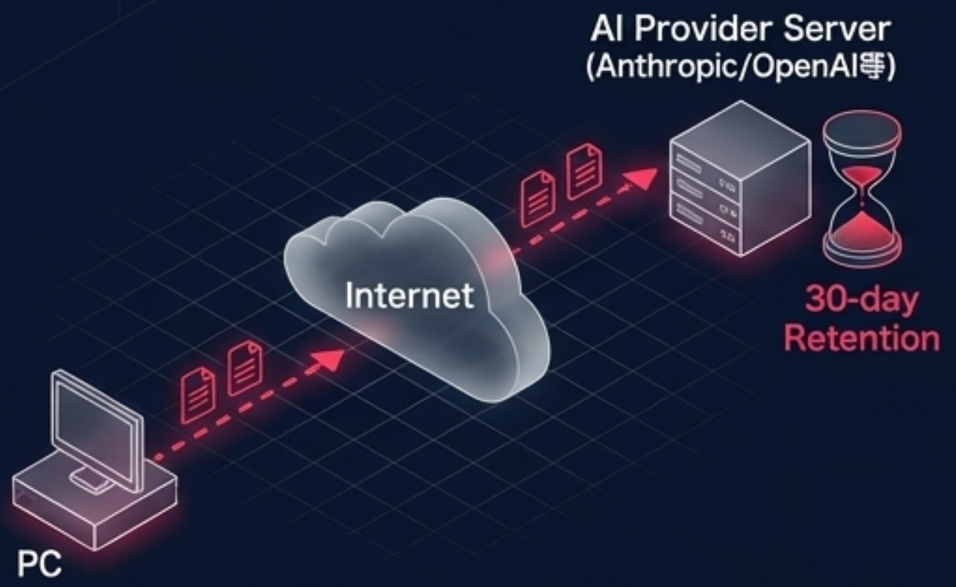
- 閉域環境での利用: AWS Bedrock / Google Vertex AI経由での利用。PrivateLinkを利用した自社VPC内完結構成。外部プロバイダ側にデータが一切渡らない。

防壁・境界 (Governance & Boundaries)

- Technical Boundary（技術的境界): Zero Data Retention (ZDR）の適用。AI推論後、データは即時破棄され、プロバイダ側のインシデントリスクをゼロにする。
- Operational Boundary（運用的境界）: CloudTrail等を利用した全操作の完全な監査証跡（180日以上の保存）。Compliance APIによる定期レビュー。ISO 42001やFISC基準の要件を満たす強力なアカウンタビリティ。

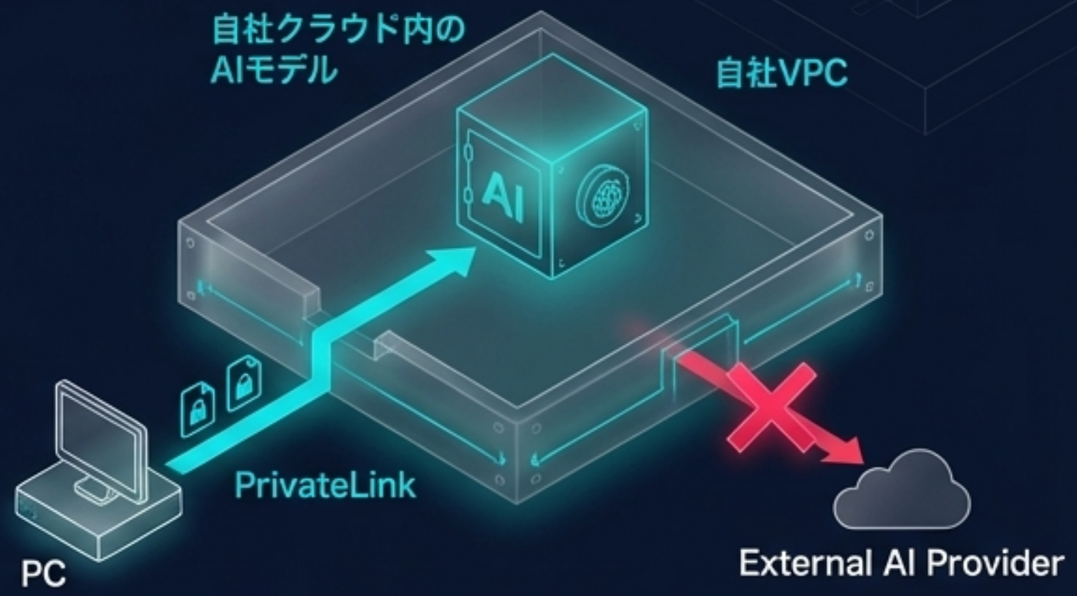
Infrastructure Deep Dive: データはどこへ行くのか？

Direct API (Level 1/2)



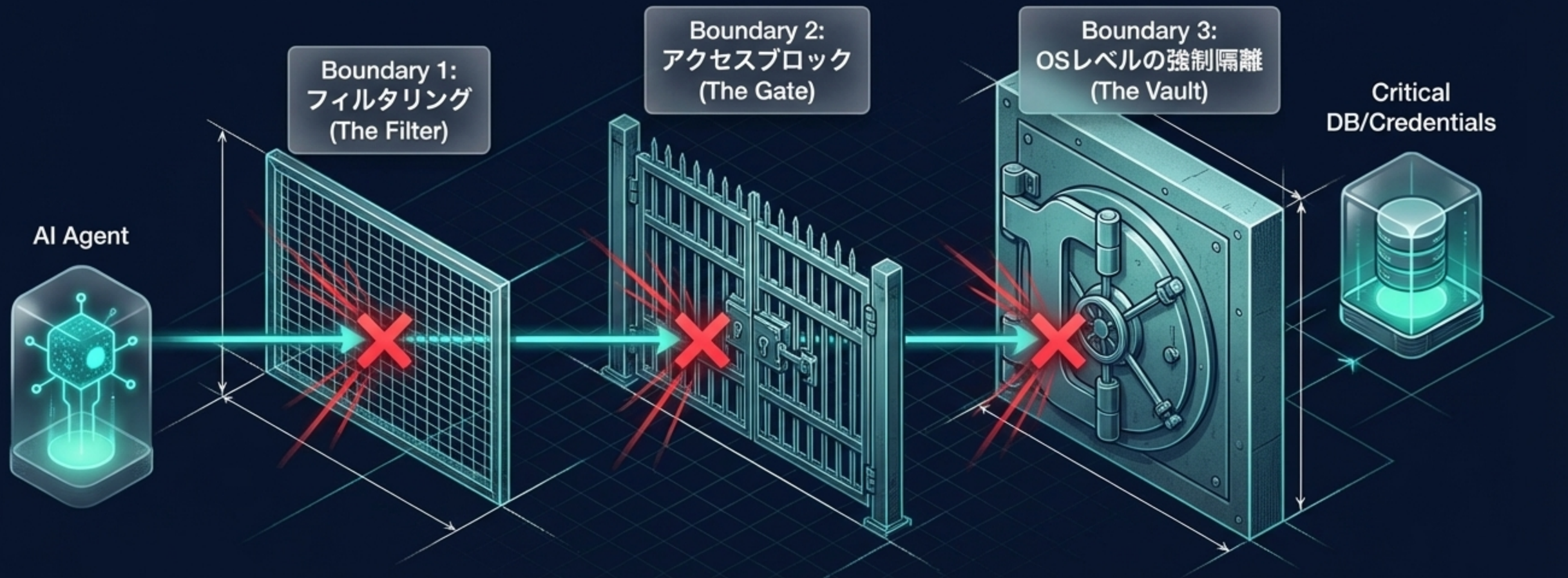
- Flow: 自社PC → インターネット → AIプロバイダ (Anthropic/OpenAI等)
- Risk Profile: 学習には使われない（契約による保証）。しかし、データは最大30日間プロバイダのサーバーに保持される。プロバイダ側が侵害された場合の理論的リスクが残存。

AWS Bedrock / GCP Vertex (Level 3)



- Flow: 自社PC → PrivateLink → 自社クラウド内のAIモデル
- Risk Profile: AIプロバイダにデータは一切送信されない。データは自社クラウドインフラ内で完結し、推論後即時破棄 (ZDR)。「外部のインシデント」によるコード流出リスクを完全に排除。

The Secure Engine



- .claudeignore 構文を用いた機密DBや認証情報へのアクセス除外（個人情報保護法の安全管理措置に対応）。
- settings.json の deny ルール実装。危険なシェルコマンド（curl, wget, ssh等）の実行を未然にブロック。
- サンドボックスモード（Seatbelt / bubblewrap）の有効化。エージェントの暴走やプロンプトインジェクションが発生しても、ファイルシステムやネットワーク境界を越えられない最終防衛線。

Designing Operational Boundaries: 組織を守る「運用」の設計

Mechanism 1: パーミッションと承認 (Human-in-the-Loop)

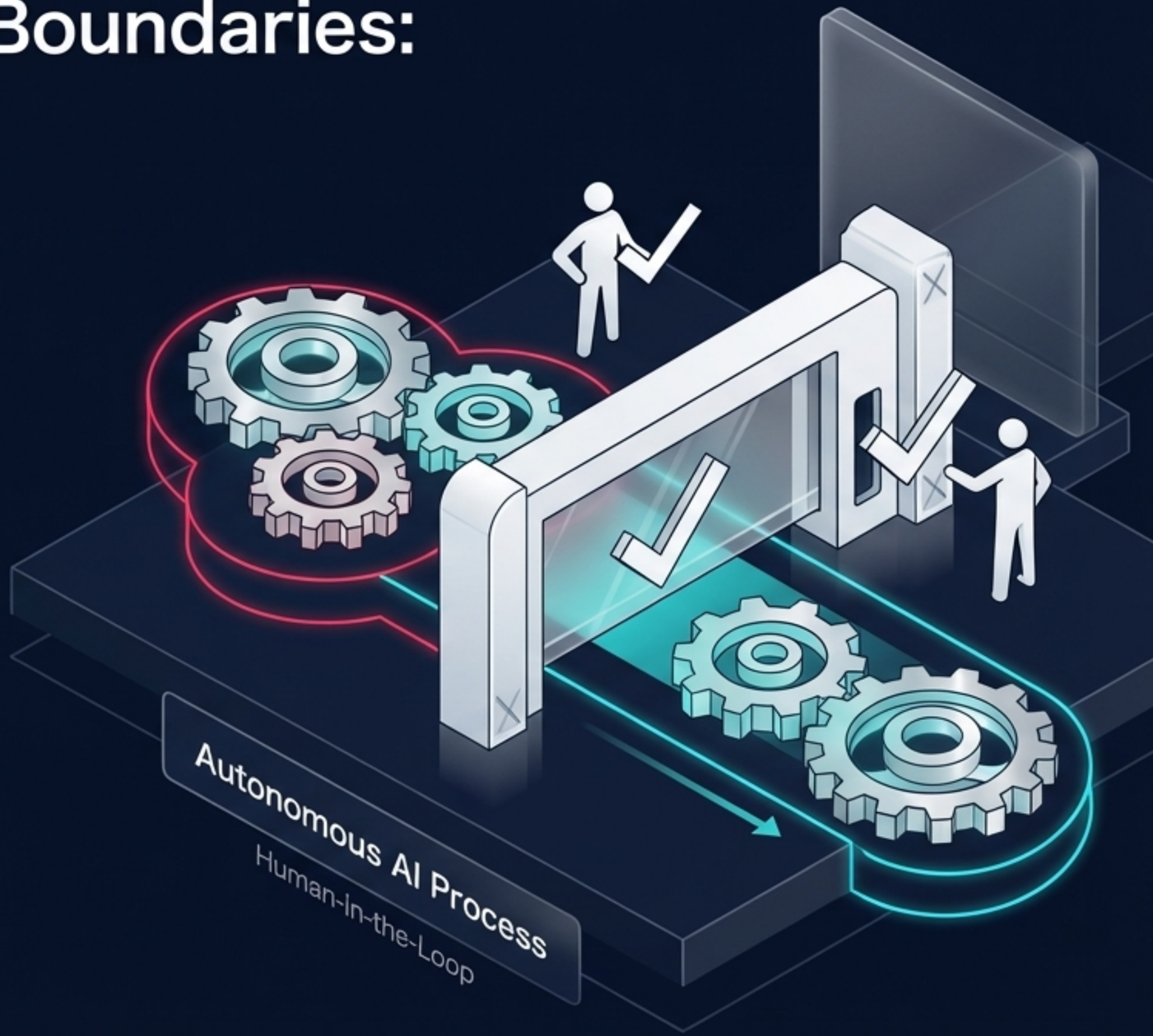
・AIによるコード変更や重要なコマンド実行時に、必ず「明示的なユーザー承認ダイアログ（Ask / Allow / Deny）」を挟む運用。AIの暴走を直前で遮断する。

Mechanism 2: 追跡と検証 (Full Traceability)

・「誰が・いつ・何を」AIに指示したかの完全な監査証跡の確保。フローセッションとCI/CDログの連携。

Mechanism 3: ツールのホワイトリスト化 (Sanctioned Tooling)

・セキュリティ部門が許可したAIツール・モデル・プロンプトのみを使用可能にし、シャドーAIを根絶するガバナンス運用。



The 3-Tier AI Governance Matrix: セキュリティレベル別導入ガイド

	推奨インフラ	データ保持	認証・権限	技術的境界	監査・ログ
Level 1 (Startup)	Direct API	30日	なし	.ignore	なし
Level 2 (Mid-Size)	組織向けSaaS	30日	SSO連携	サンドボックス	基本ログ
Level 3 (Enterprise / Regulated)	閉域環境 (Bedrock/Vertex)	ZDR (実質0日)	SSO + IAM最小権限	サンドボックス + PrivateLink	180日監査ログ (CloudTrail)

自社の扱うデータ（公開情報か、社内機密か、要配慮個人情報か）に基づき、ターゲットとなるLevelを決定する。

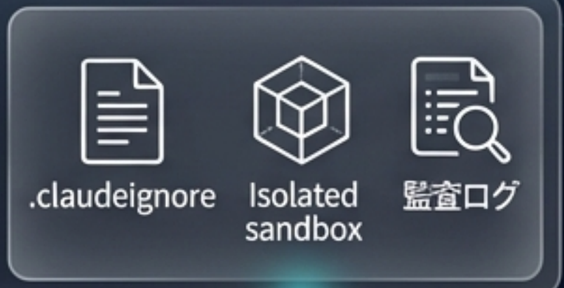
Regulatory Alignment: 日本の法規制・ガイドラインとの完全な整合

個人情報保護法対応 (PPC)

- 学習不使用の契約保証: Level 1〜3すべてで対応済み（第三者提供の委託整理）。
- 第三者提供制限の回避: Level 3 (Bedrock経由) なら外部へデータが渡らず、第三者提供自体が発生しない。

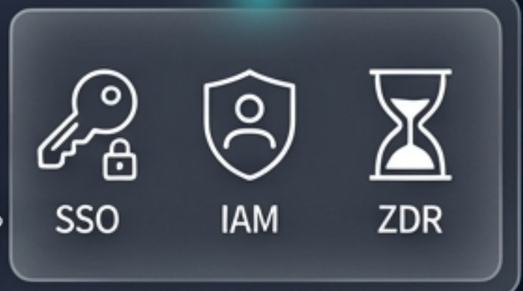
AI事業者ガイドライン（指針4/5/6/7）

- プライバシー保護、セキュリティ、透明性、アカウンタビリティの全項目に対し、.claudeignore、サンドボックス、監査ログで技術的にアンサーを提示可能。



金融・保険業界 (Need to Know原則 / FISC)

- Level 3構成（SSO + IAM制御 + ZDR）により、業務遂行上の必要性がある者のみにアクセスを限定し、厳格な監査基準をクリア。



Continuous AI Assurance: 継続的なAI
継続的なAIガバナンスへの進化

Phase 0: Discovery (発見)

- シャドーAIのカタログ化、未承認のAIワークロードの検出、アシュアランス負債の評価。

Phase 1: Initial (Level 1導入)

- 基本的なSBOM、手動のプロビジョニング、ポリシー設定による初期コンプライアンスの達成。

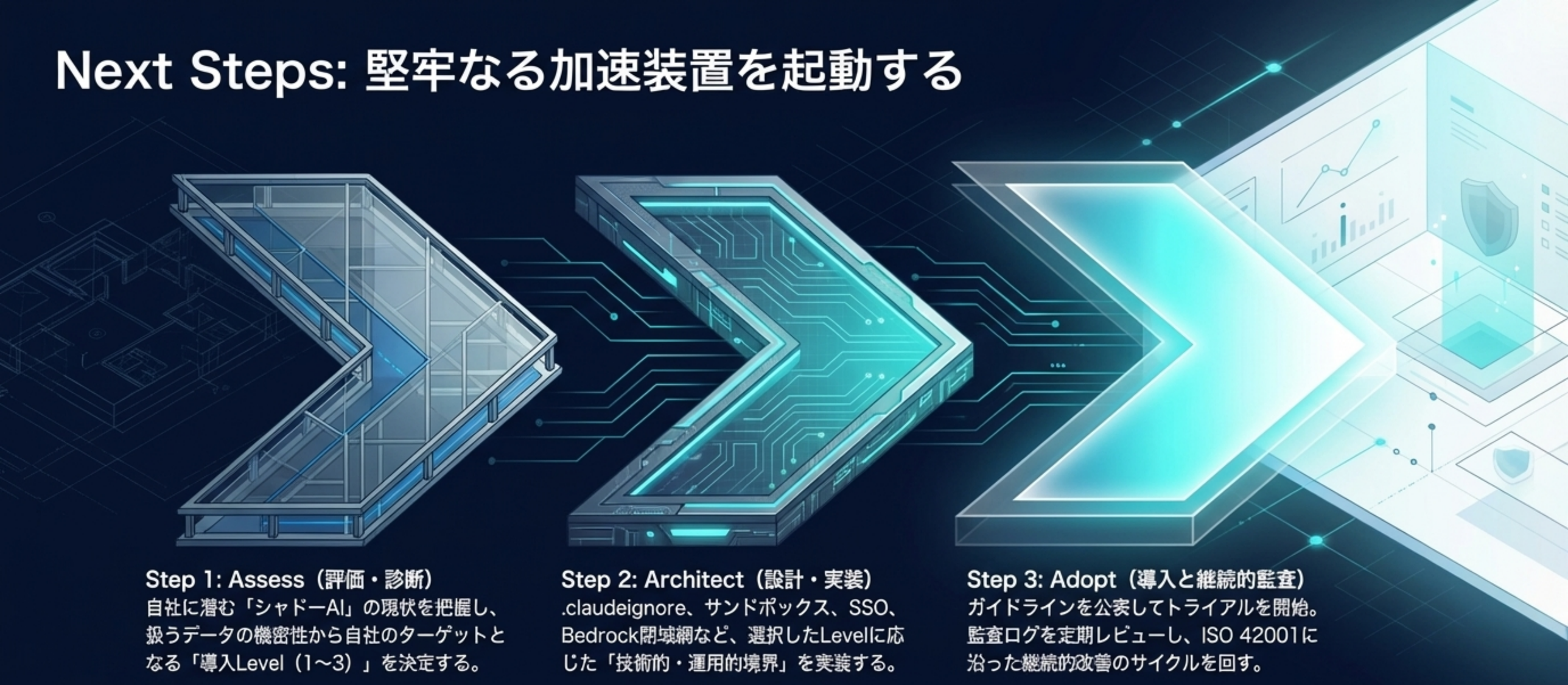
Phase 2: Managed (Level 2導入)

- 自動化された整合性チェック、SSO統合、一元化されたポリシー適用。

Phase 3: Continuous (Level 3最適化)

- ISO/IEC 42001認定への準拠、リアルタイムの異常検知と継続的なAIアシュアランス。

Next Steps: 堅牢なる加速装置を起動する



Step 1: Assess（評価・診断）
自社に潜む「シャドーAI」の現状を把握し、扱うデータの機密性から自社のターゲットとなる「導入Level（1〜3）」を決定する。

Step 2: Architect（設計・実装）
.claudeignore、サンドボックス、SSO、Bedrock閉域網など、選択したLevelに応じた「技術的・運用的境界」を実装する。

Step 3: Adopt（導入と継続的監査）
ガイドラインを公表してトライアルを開始。監査ログを定期レビューし、ISO 42001に沿った継続的改善のサイクルを回す。

「リスクがあるから使わない」のではない。「リスクを理解し、アーキテクチャで制御する」企業だけが、安全に AI の真の生産性を手にする。