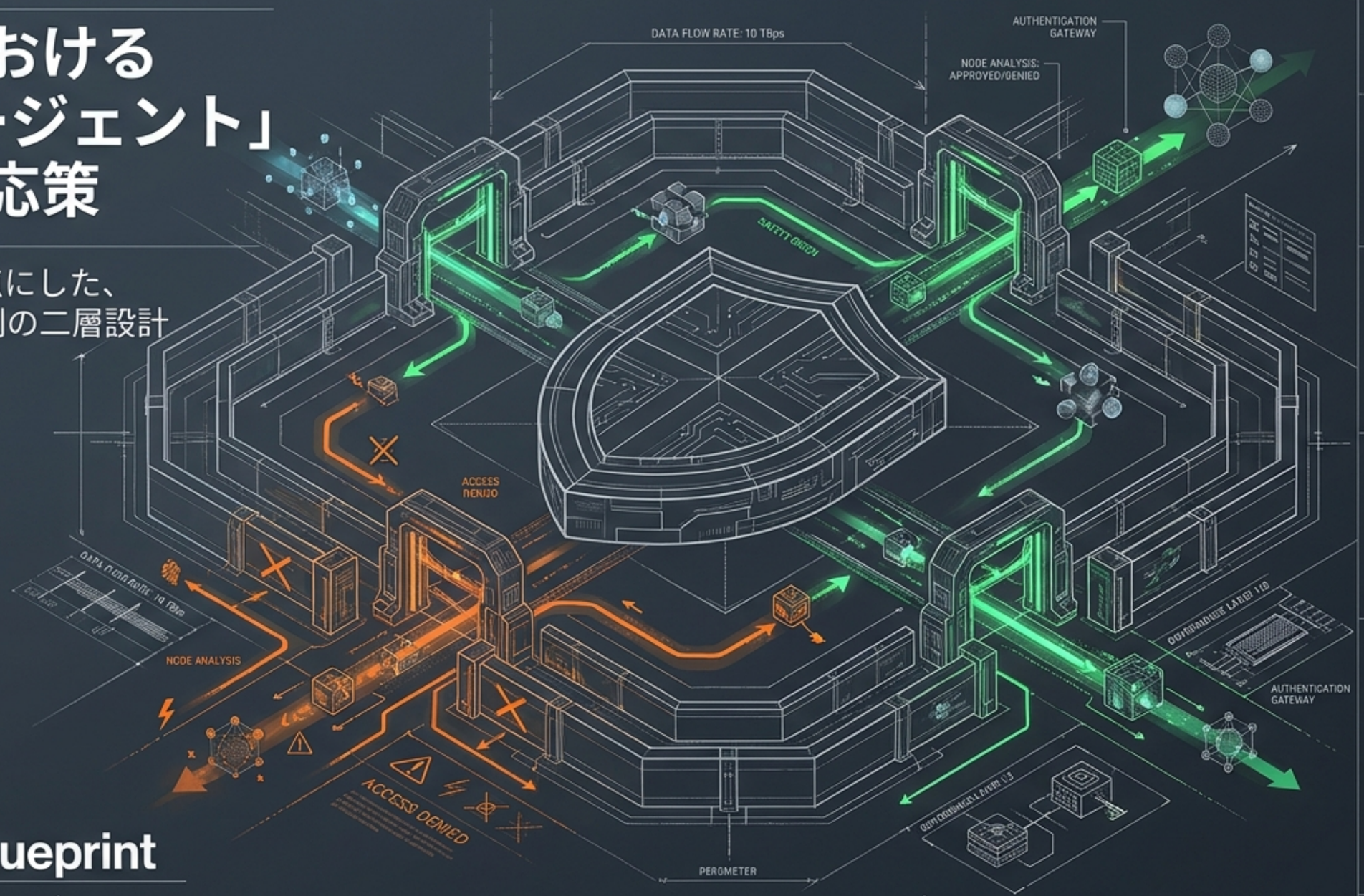
知財部門における「野良AIエージェント」問題への対応策

野良RPAの教訓を起点にした、
ガバナンスと技術統制の二層設計



The Security Blueprint

野良RPAの失敗モード

「動かなくなって困る」

業務の突発的な停止（引継ぎ欠落、ID依存）

野良AIエージェントの失敗モード

「意図せず動いて困る」

致命的な情報漏えいと特許喪失。未公開発明や出願戦略などの「極秘情報」と「自律的動作」の最悪の掛け合わせ。

発生要因（共通）：現場の効率化欲求 / IT部門のキャパシティ不足 / 開発の民主化 / 属人化		
	[RPA]	[AI Agent]
挙動 (Behavior)	決定論的（プログラミング通り）	**非決定論的**（モデル差替えで突然変化）
接続性 (Connectivity)	内部システム中心、閉じた軌跡	**外部接続性**（MCP経由で外界へ自律的書き込み）
攻撃面 (Attack Surface)	パスワード・IDの漏洩	**プロンプトインジェクション**（入力自体が攻撃ベクトル）
責任追跡 (Traceability)	単一の実行ログ	**エージェント連鎖**（副エージェントへのタスク委譲による追跡困難）

Lethal Trifecta（致命的三条件）と知財ワークフロー

条件1：
プライベートデータ
へのアクセス

条件2：
信頼できないコ
ンテンツへの露出

条件3：
外部通信能力

Data Exfiltration
(データ流出)

知財業務：
社内発明提案書の
読み込み

知財業務：
外部特許文献
APIの取得

知財業務：
外部LLMによる処理と
自動メール返信

知財部門の「典型的な業務ワークフロー」自体が、AIエージェントに
におけるデータ流出の構造的脆弱性を完璧に満たしている。

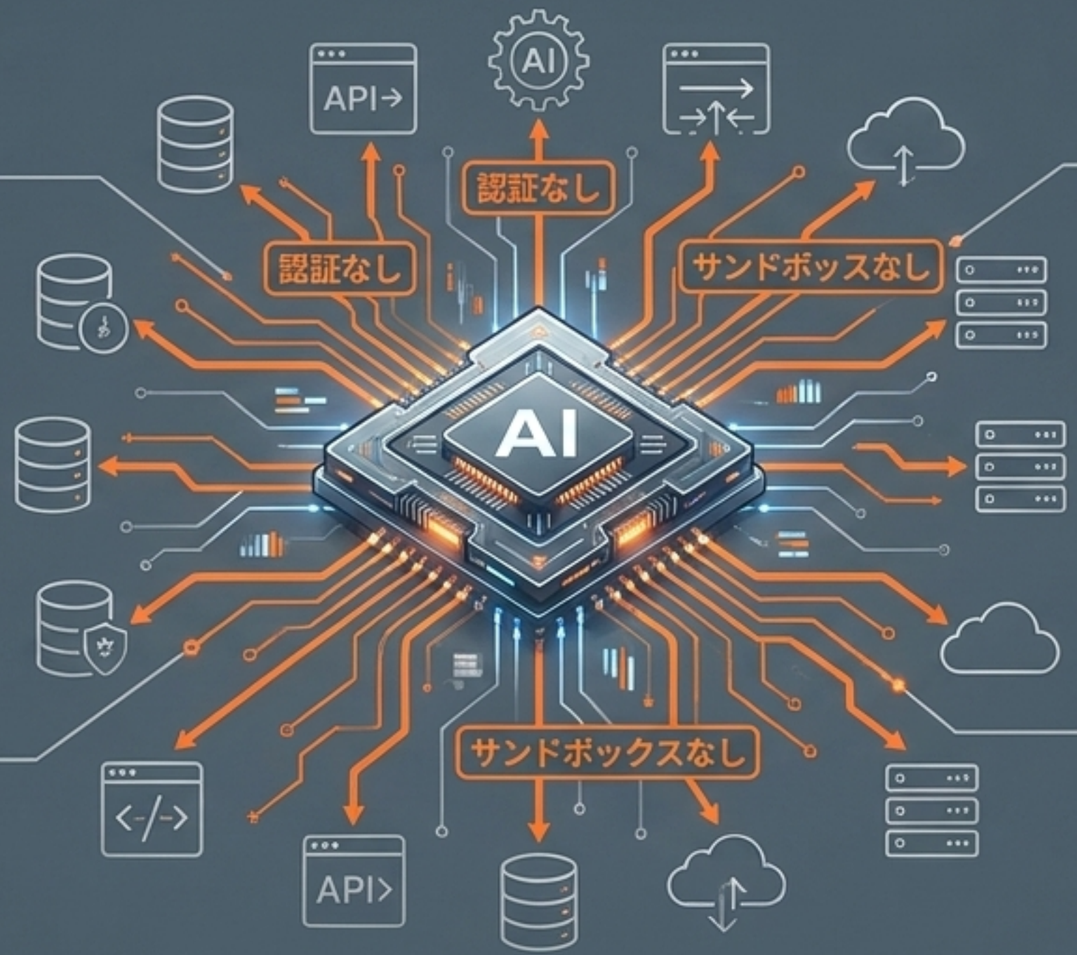
2025年最大の攻撃面：MCP (Model Context Protocol)

規模の爆発

2025年中にGitHub上のMCPサーバは13,000件超へ。セキュリティ監視は事実上追いつかない。

OWASP第1位

Agentic Applications 2026の最大脅威は「Agent Goal Hijacking」。



設計上の脆弱性

MCP仕様は通信機構のみ。認証・認可・サンドボックスを「強制しない」構造的欠陥。

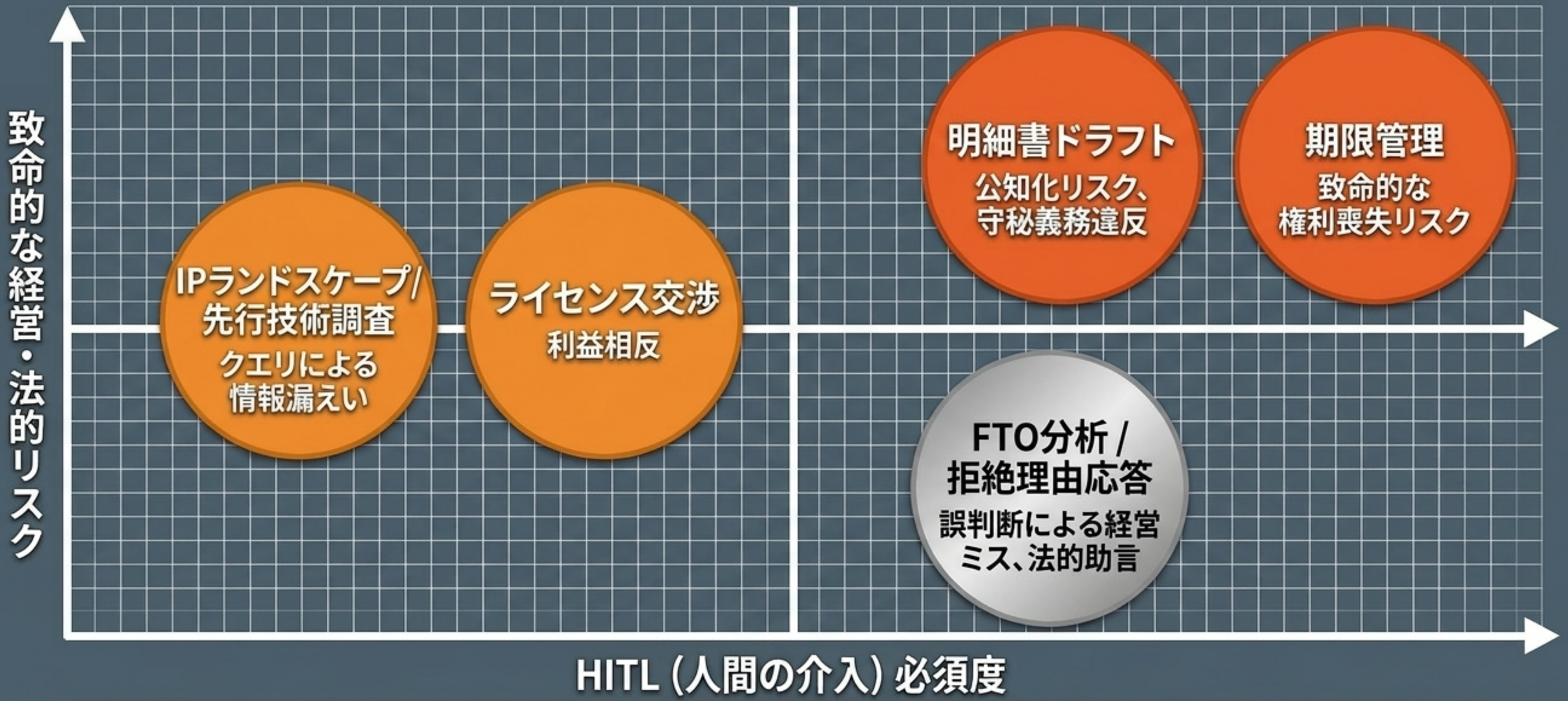
実被害の顕在化

Asana MCPのテナント間データ露出、Microsoft 365 Copilotの脆弱性、SQLインジェクションの確認。

2025年、一斉に整った「制度的フレームワーク」

もはや野良AIの黙認は法令・規範違反に直結する。

知財ユースケース別のリスクと推奨制御



業務ごとにリスク特性が全く異なるため、「一律禁止」も「一律許可」も成立しない。

なぜ「全面禁止」は失敗するのか

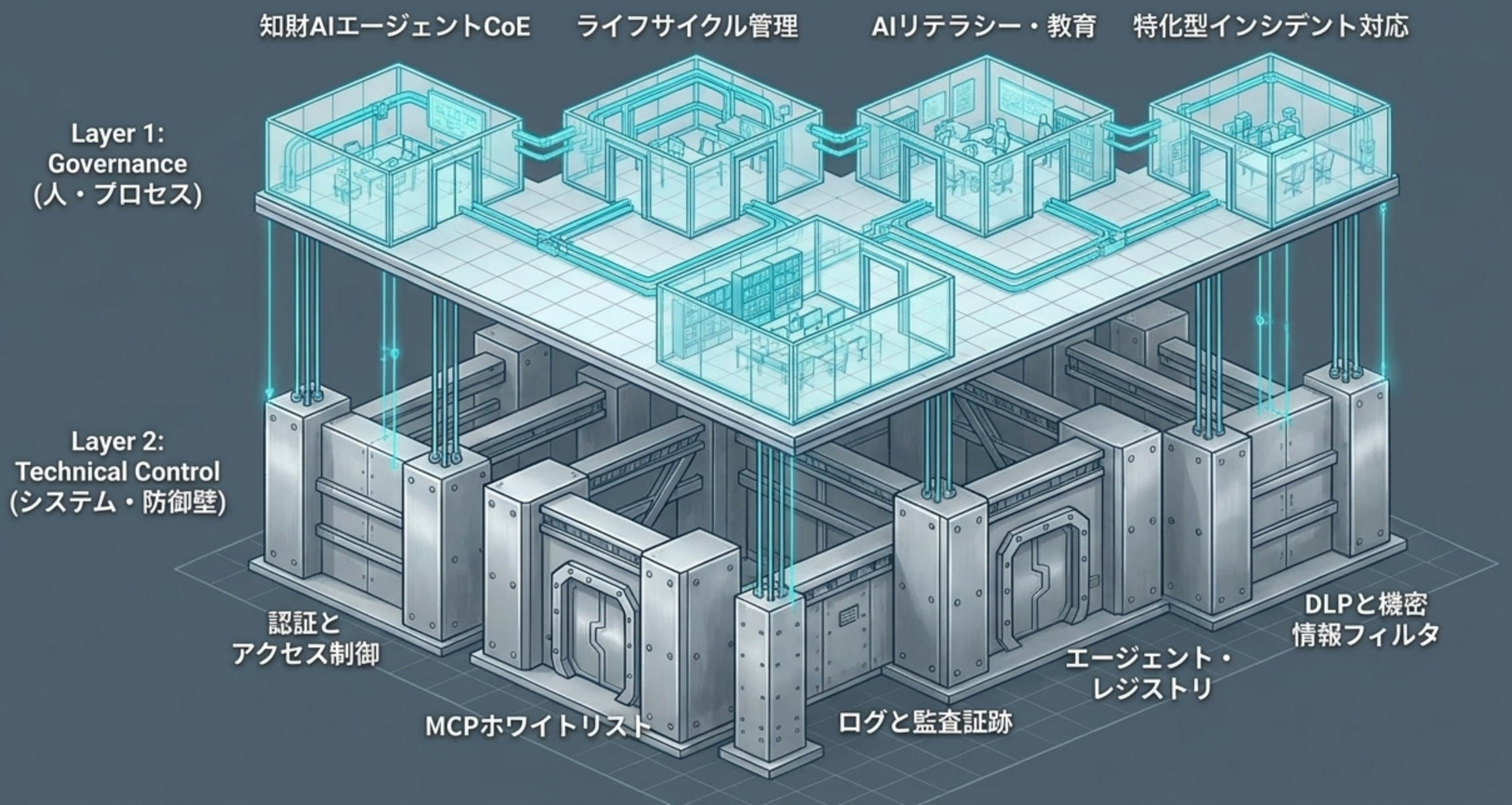
インシデント追加コスト

平均 19万米ドル（約2,900万円）

ソリューションの三点セット

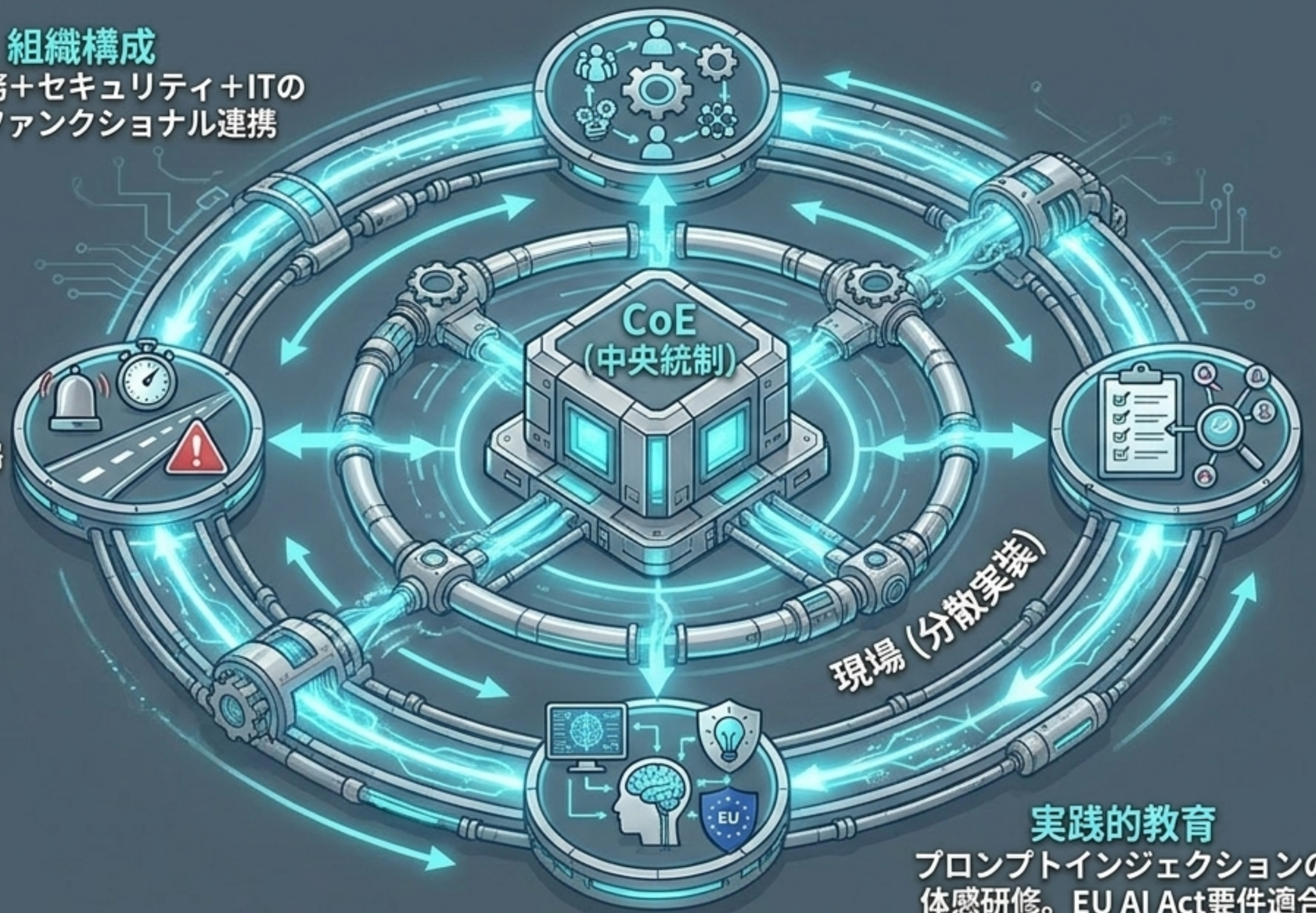
許可リスト方式 （ホワイトリストの運用）	安全な代替環境 （社内RAG・閉域API）	可否マトリクス （情報分類別の整備）

ガバナンスと技術統制の二層設計



Tier 1: 知財AIエージェント CoEとライフサイクル

組織構成
知財＋法務＋セキュリティ＋ITの
クロスファンクショナル連携

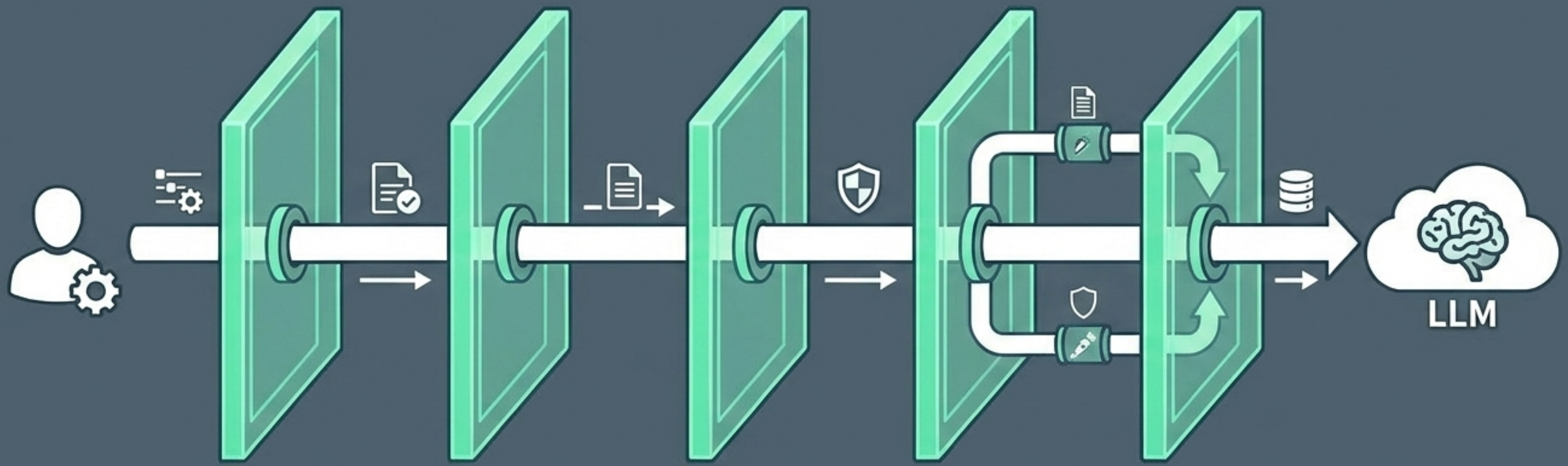


インシデント対応
特許出願期限直前用の
別エスカレーション経路
(ファストトラック)

独自アセスメント
NIST AI 600-1 ＋
知財固有3リスク
(特許要件喪失、守秘
義務、職務発明整合)

実践的教育
プロンプトインジェクションの
体感研修。EU AI Act要件適合

Tier 2: 自動化された技術統制と防御壁



1. Identity Control

人間用IDの貸与禁止。APIキーのボルト管理と即時Revoke。

2. Agent Registry

目的、所有者の中央台帳。自動化された定期健康診断。

3. MCP Control

GitHub公開サーバーの直接利用禁止。組織承認済みのホワイトリストのみ許可。

4. DLP & Filtering

出願戦略、未公開発明の自動マスキング。CASBによる遮断。

5. Observability

OpenTelemetry標準でのLLM推論・MCP通信ログのSIEM出力。

実装ロードマップ (Months 0–6)

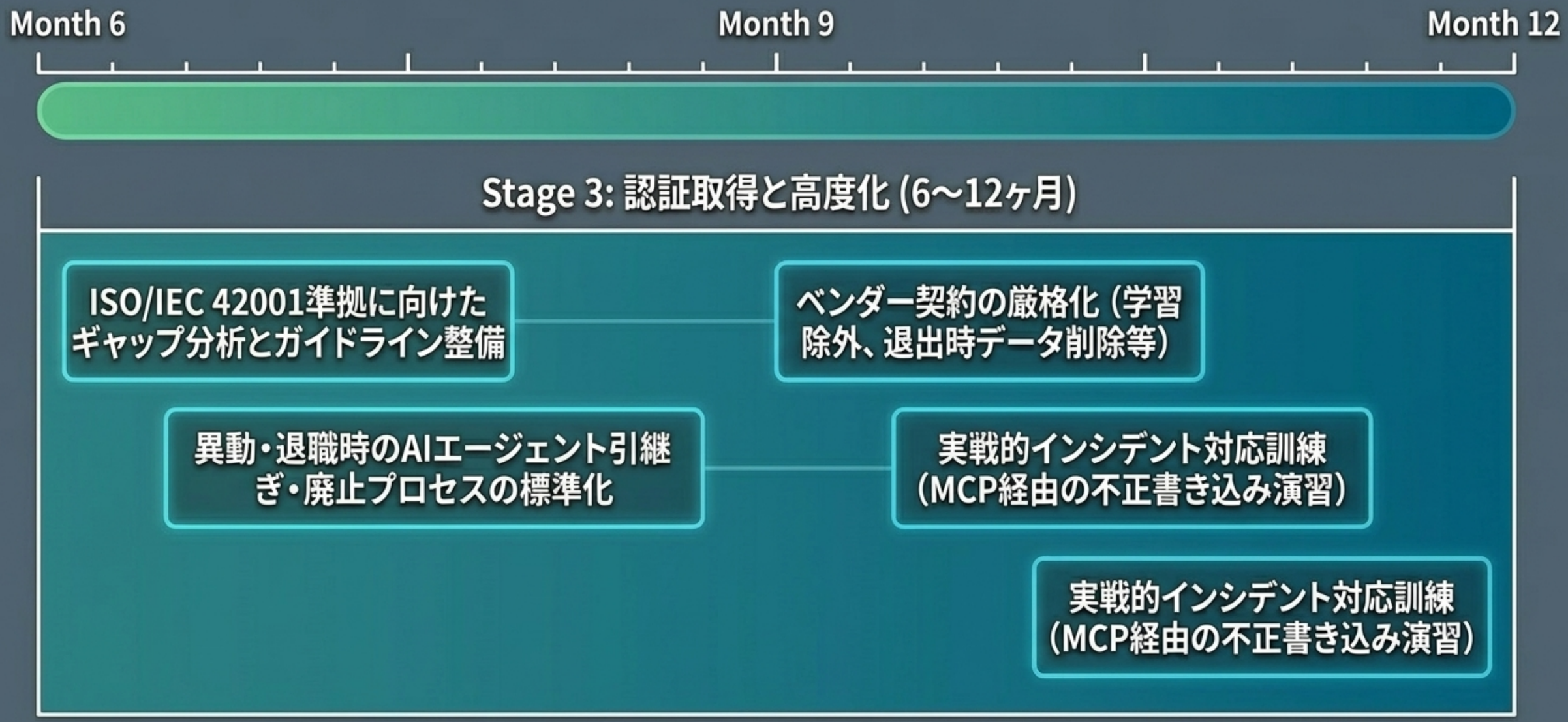
Stage 1: 出血を止める (即時〜3ヶ月)

- CASBとPCログによる野良AIエージェントの強制棚卸し
- トップ通達: 未公開発明と出願情報は秘密保持義務環境のみに入力
- 個人のAPIキー、スプレッドシート埋め込みキーの即時Revoke

Stage 2: CoE設立と標準化 (3〜6ヶ月)

- 知財AI CoEの発足と、利用可否マトリクスの策定
- 監査ログ基盤とMCPホワイトリストの初期導入
- 弁理士・知財担当者向けのハルシネーション体験型教育

実装ロードマップ (Months 6–12)



The New Standard

AIエージェントの圧倒的な恩恵を取り込みつつ、
知財実務の絶対的信頼性を守る。

その両立を可能にするのは「現場任せ」でも「無意味な禁止令」でもない。

CoEを中核とする、ガバナンスと技術統制の
「絶え間ない継続的運用（Continuous Governance）」である。

法令、指針、国際規格。2025年中に出揃った制度的圧力をテコとし、
今すぐ「最低限のベースライン」の構築に着手せよ。