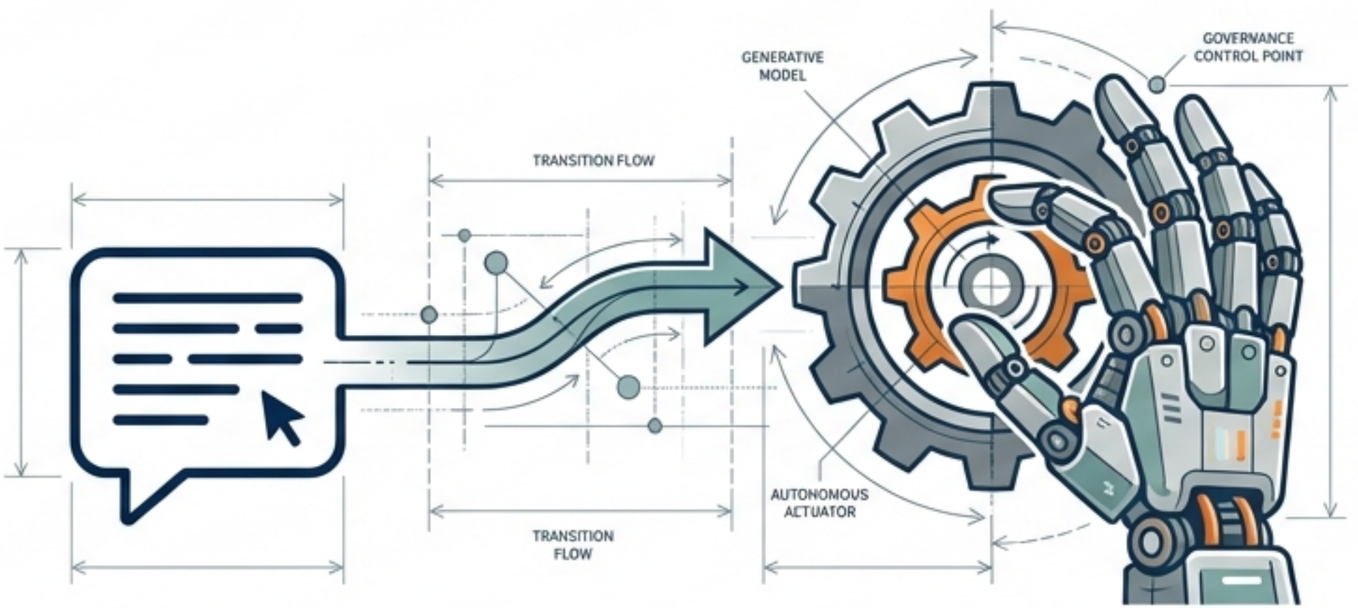
日本政府AI事業者ガイドライン改定案の分析

「生成」から「行動」へ：AIエージェント・フィジカルAI時代のガバナンスと統制設計



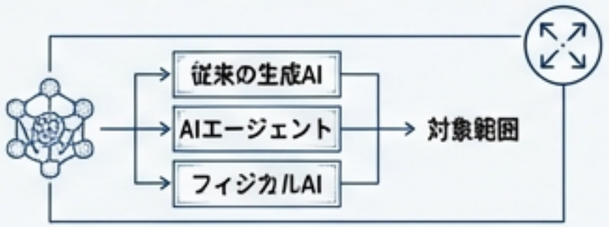
GENERATIVE AI
(生成AI)

AGENCY & ACTION
(行動・エージェント)

エグゼクティブサマリ：静的AIから動的AIへのガバナンス転換

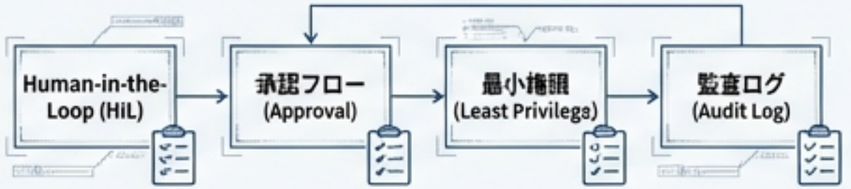
適用範囲の拡大

従来の「**生成AI**」に加え、自律的にタスクを遂行する「**AIエージェント**」と、現実世界に作用する「**フィジカルAI**」が明確な対象となる。



システム要件化

「人間の判断必須（Human-in-the-Loop）」はスローガンではなく、**承認フロー、最小権限、監査ログ**といった具体的エンジニアリング要件として定義される。

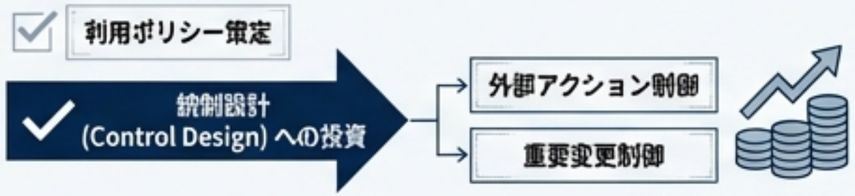


法的実効性

ガイドライン自体は「**ソフトロー**」だが、契約・調達における「**説明責任の共通言語**」となり、事故時の「**標準的な安全配慮義務**」の基準として機能する。

求められるアクション

利用ポリシーの策定だけでなく、「**外部アクション**」や「**重要変更**」を制御するための**統制設計（Control Design）**への投資が不可欠となる。



パラダイムシフト：コンテンツリスクから「行動リスク」へ

従来のガイドライン (Static)

対象: Chatbots, Content Generators
性質: 思考・生成 (Thinking)
主要リスク:
- 著作権侵害
- 偽情報
- バイアス

統制: 人間による出力確認

改定案 (Kinetic)

対象: AI Agents, Autonomous Robots
性質: 行動・実行 (Doing)
主要リスク:
- 勝手な資金移動
- システム停止
- 物理的危害
- プライバシー侵害

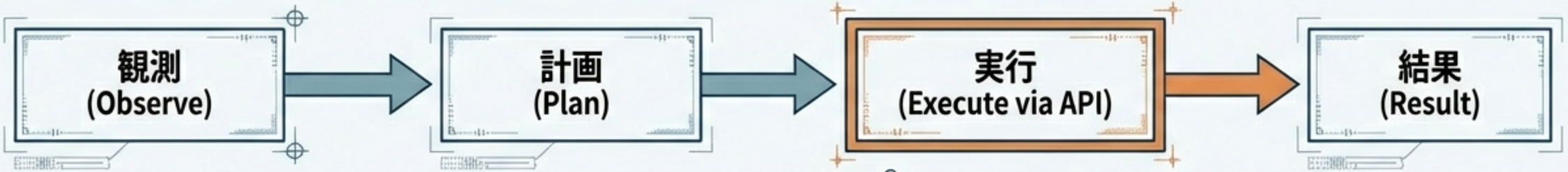
統制: アーキテクチャによるガードレール
(Permissions, Kill switches)

Insight:「生成する（Generate）」ことではなく「実行する（Execute）」ことにガバナンスの重心が移動している。

新たな規律対象の定義：AIエージェントとフィジカルAI

1. AI Agent（AIエージェント）

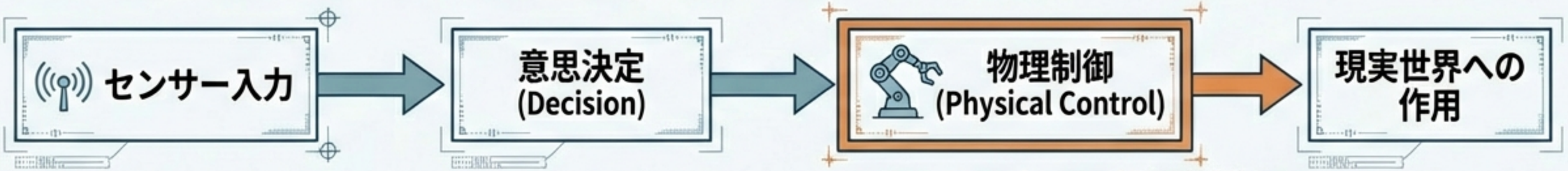
目標達成のために計画を立て、ツール（API）を操作して自律実行するシステム。



デジタル空間での「自律的な外部アクション」

2. Physical AI (フィジカルAI / ロボAI)

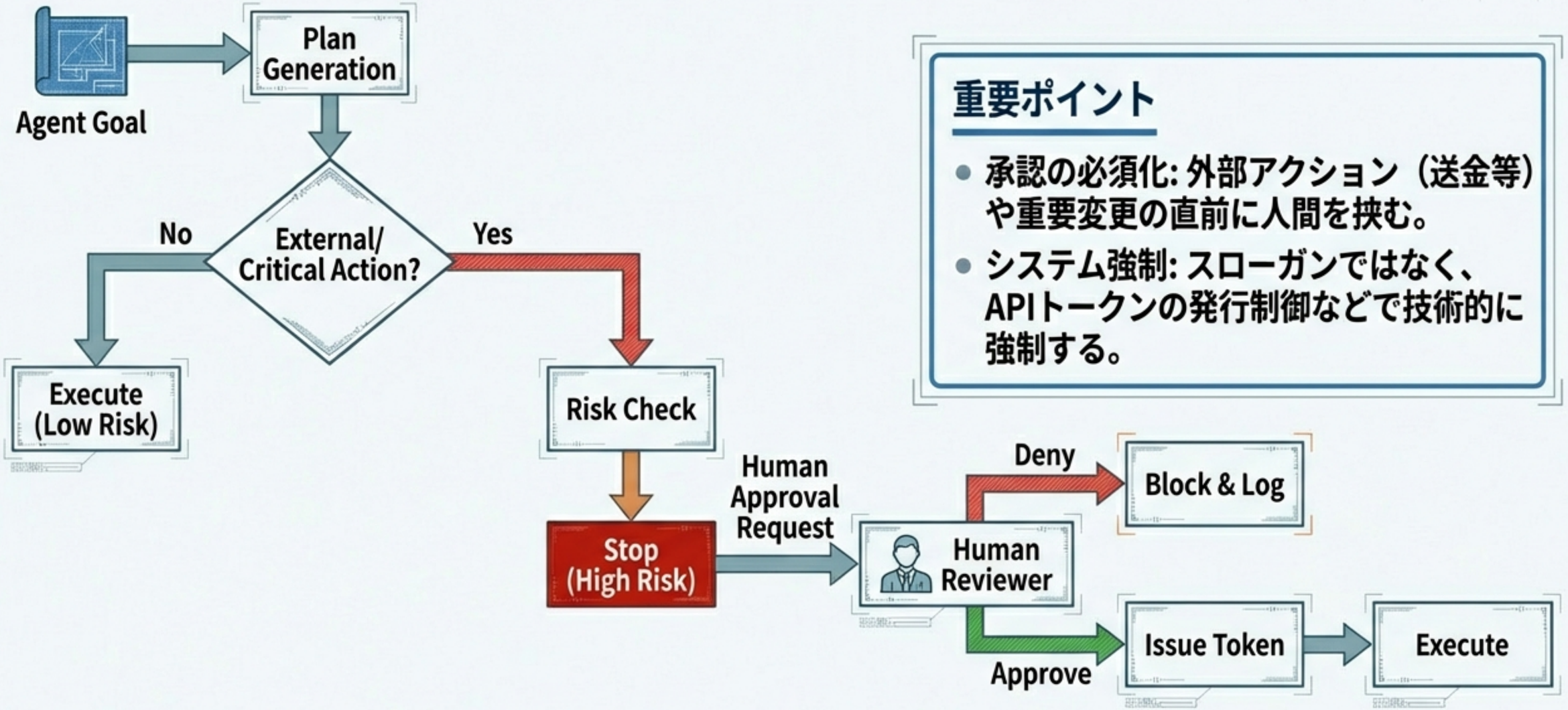
センサー入力に基づき、物理デバイスの制御を行うシステム。



ソフトウェアの誤作動が直ちに
「物理的安全」を脅かす

※これらの定義はワーキング定義であり、最終版での境界事例（RPA+LLM等）の確認が必要。

統制アーキテクチャ：「Human-in-the-Loop」の実装モデル



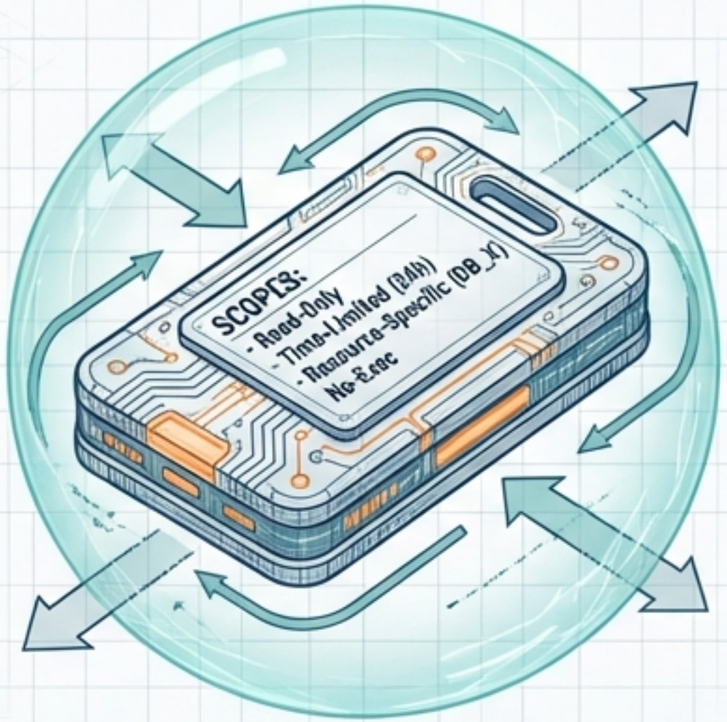
重要ポイント

- 承認の必須化: 外部アクション（送金等）や重要変更の直前に人間を挟む。
- システム強制: スローガンではなく、APIトークンの発行制御などで技術的に強制する。

※これらの定義はワーキング定義であり、最終版での境界事例（RPA+LLM等）の確認が必要。

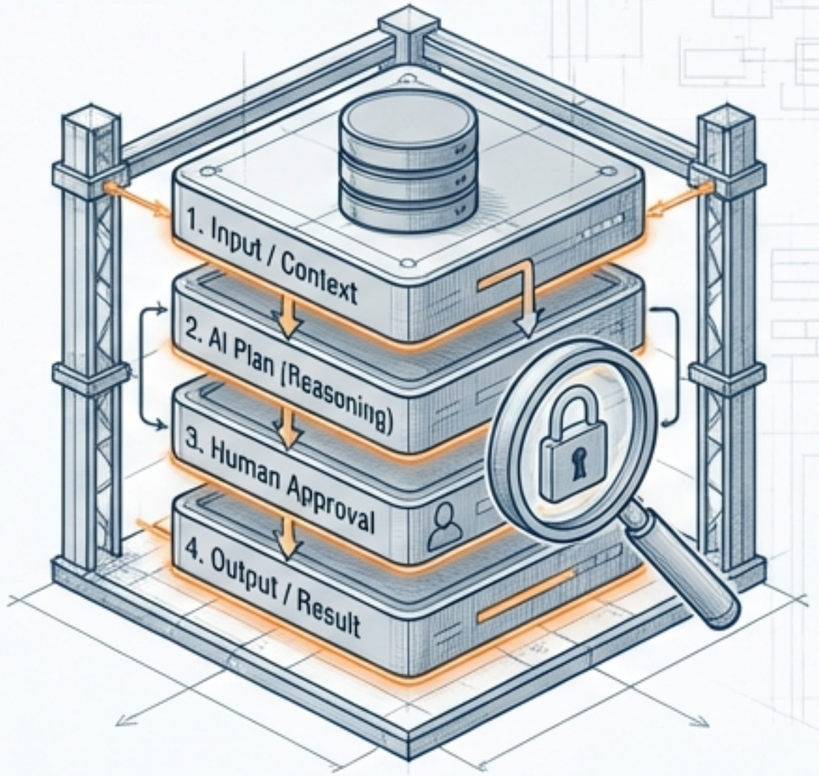
エンジニアリング統制：最小権限と監査ログ

Least Privilege (最小権限の原則)



- **概念:** AIには「そのタスクに必要な最小限の権限・期間・範囲」のみを付与する。
- **実装:** スコープ制限、有効期限付きトークン。
- **目的:** 被害上限（Blast Radius）を局所化するため。

Audit Logs (監査ログと説明責任)

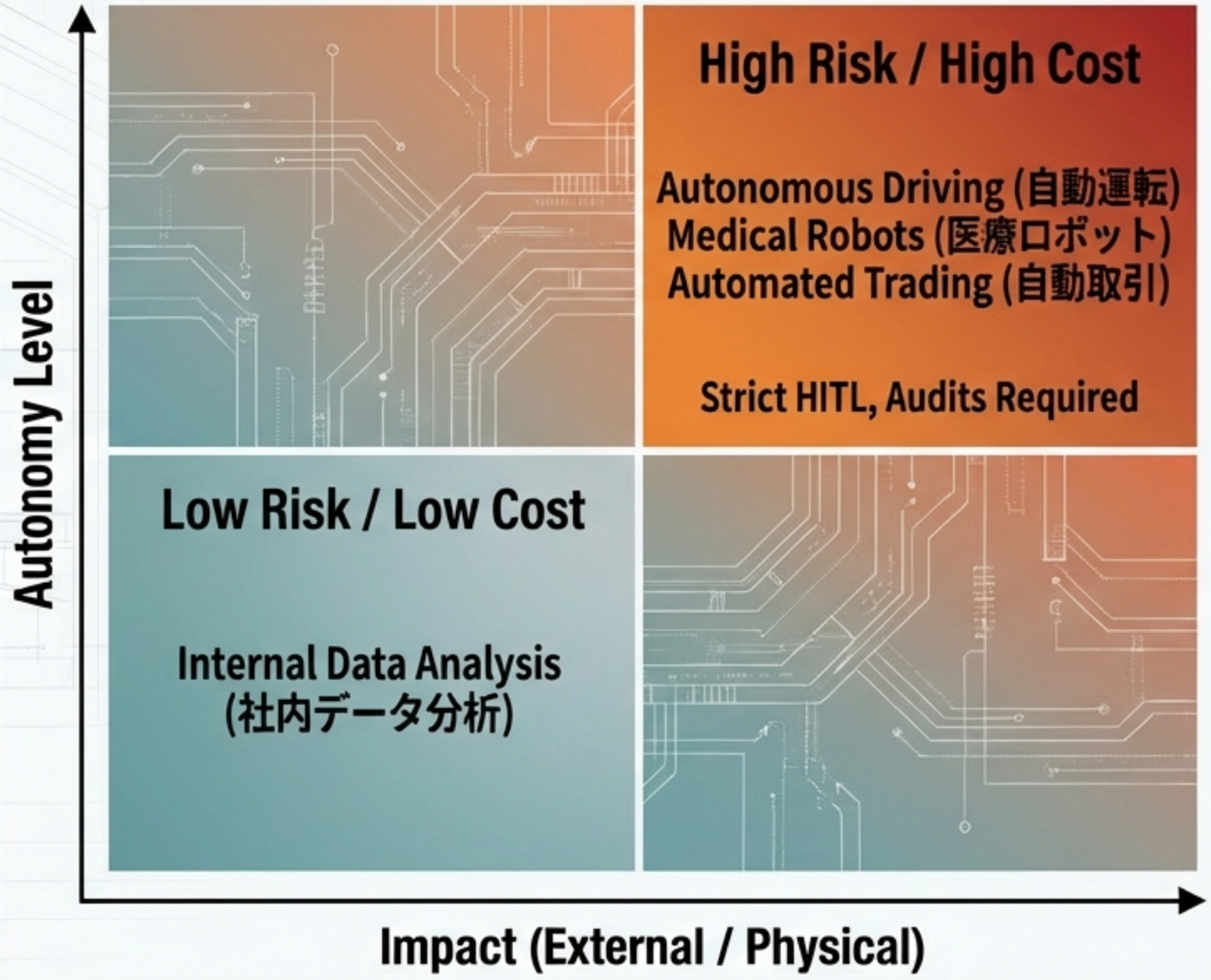


- **要件:** 「なぜその行動をとったか」を事後追跡できる完全なログ。
- **リスク:** ログ自体が機微情報を含むため、暗号化とアクセス制御が必要。

STRATEGY CONSULTING GROUP

2024年10月26日

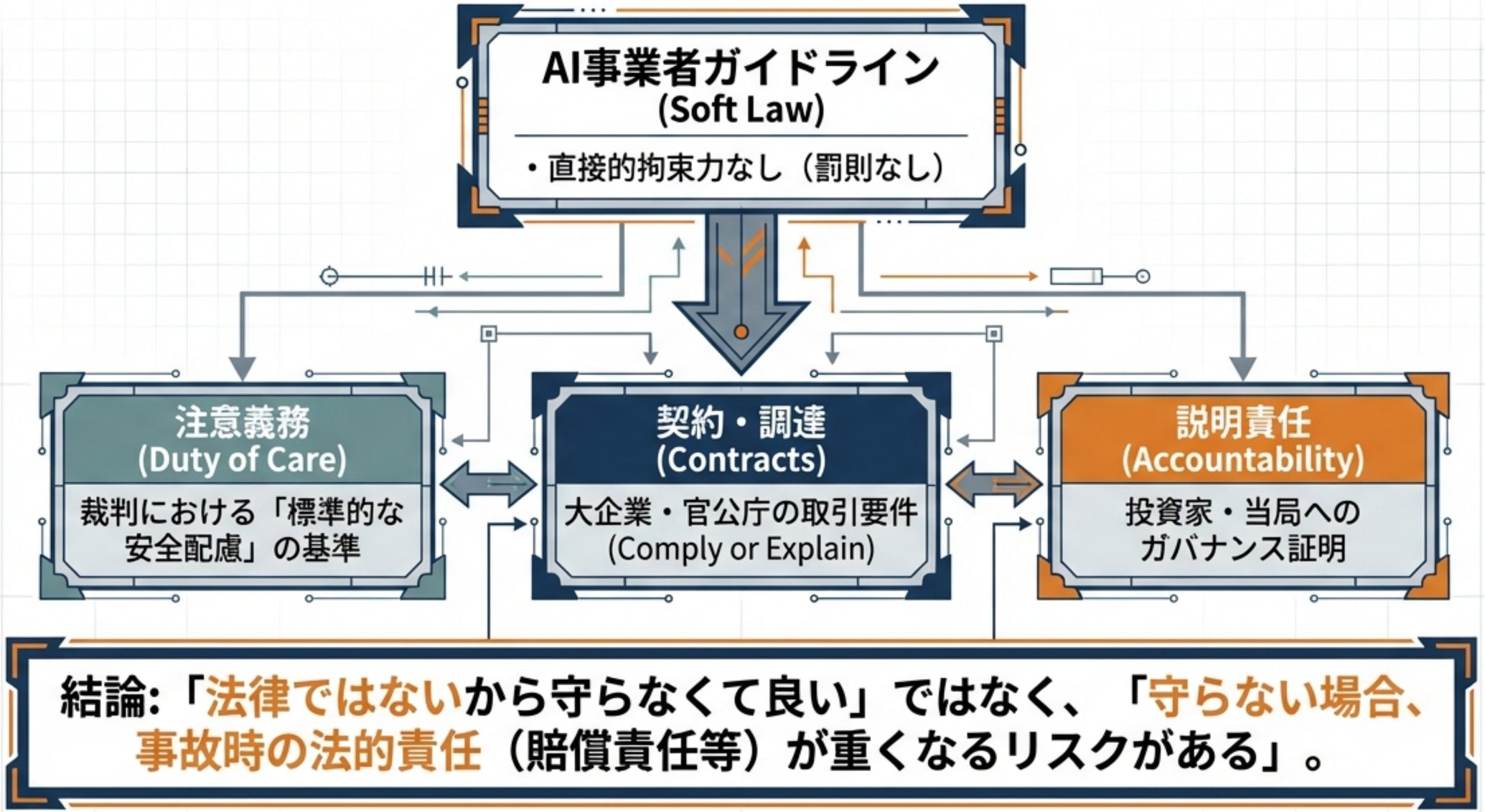
リスクベースアプローチ：コストと統制強度のヒートマップ



Cost Drivers（コスト要因）

- 資産性: 資金・個人情報・制御系の重要度
- 自律度: 連続行動・マルチエージェント
- 物理性: 安全工学（フェイルセーフ）の実装

法的・実務的意味合い：ソフトローとしての実効性



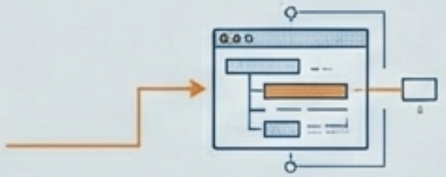
国際比較：世界の規制ランドスケープにおける日本の立ち位置

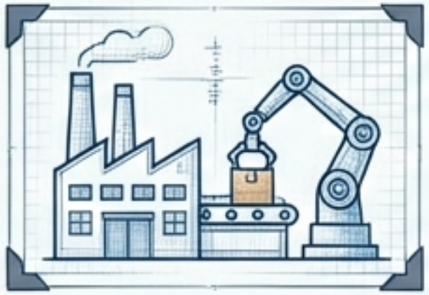
EU (Hard Law)	US (Voluntary)	Japan (Hybrid)
• Framework: EU AI Act • 性質: 法的義務・高額罰金 • 焦点: 厳格な製品安全・市場アクセス	• Framework: NIST AI RMF • 性質: 任意の枠組み • 焦点: 柔軟性とイノベーション	• Framework: AI事業者ガイドライン • 性質: ソフトロー (事実上の標準) • 戦略的価値: EU規制と米国流の間の「ブリッジ (Cross-walk)」として機能

業種別インパクト：高リスクユースケースの具体例

Finance (金融)

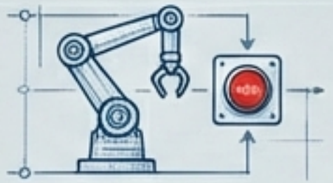
ユースケース: 融資審査、自動資産運用
リスク: 公平性、不正送金
要件: 決定過程の完全なログと「説明可能性」





Manufacturing (製造・物流)

ユースケース: 工場内AGV、協働ロボット
リスク: 労働安全（接触事故）、OT攻撃
要件: 物理的な「緊急停止（Kill Switch）」

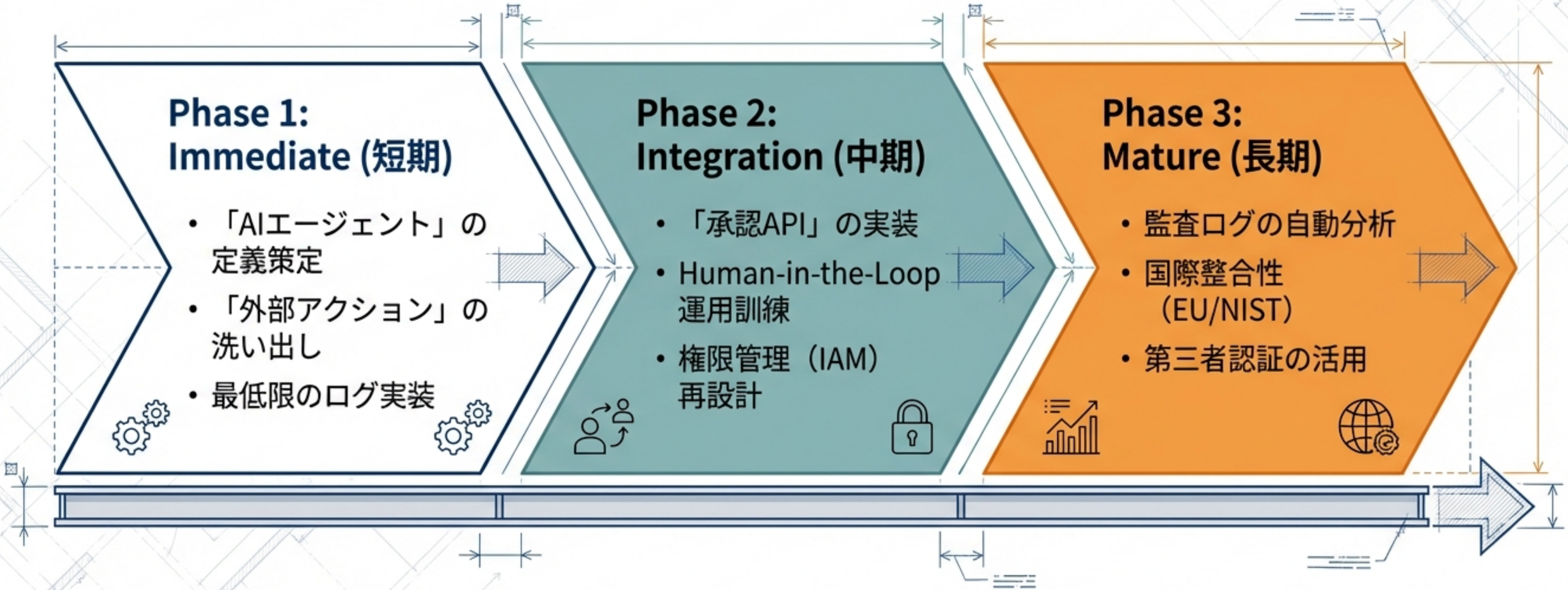


Healthcare (医療)

ユースケース: 診断支援エージェント
リスク: 誤診、プライバシー漏洩
要件: 専門家による監督（Human Oversight）

Key Metric: Blast Radius (被害半径) ——
AIが持つ「実行権限」の範囲がリスク尺度となる。

実装ロードマップ：組織が取るべき段階的アプローチ



STRATEGY CONSULTING GROUP

2024年10月26日

意思決定者のためのチェックリストと結論

Readiness Checklist

- [ ] 「外部アクション」をすべてリストアップできているか？
- [ ] AIの判断プロセスを追跡できるログがあるか？
- [ ] 「緊急停止手段（Kill Switch）」は確保されているか？
- [ ] 承認プロセスは形骸化していないか？

Shift from ‘Adoption’ to ‘Control Design’

「AIの導入」から「統制の設計」へ。
確実なガバナンスの実証は、コンプライアンスコストではなく、信頼を獲得するための最大の競争優位となる。