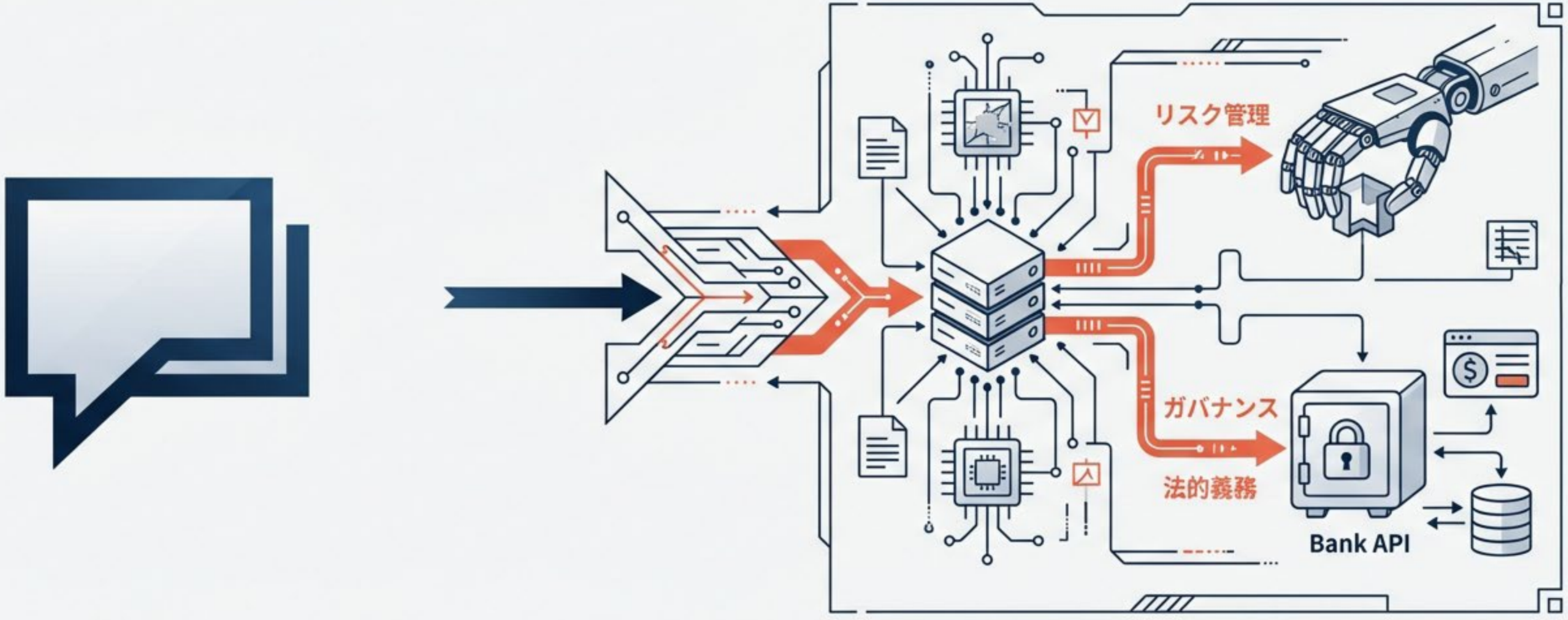
AI事業者ガイドライン改定案（第1.2版）徹底解剖

生成AIから自律型エージェントへ ― 「人間中心」の統制構造と攻めのガバナンス



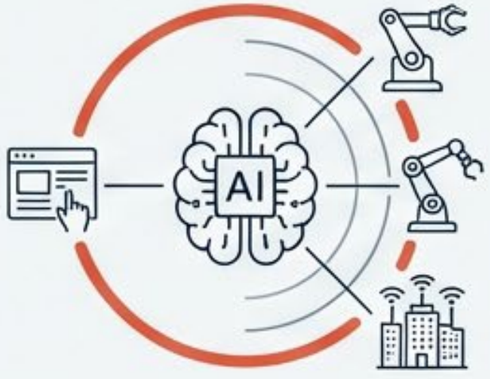
2026年改定対応・経営層向け戦略資料

エグゼクティブサマリー：2026年改定の衝撃

AIのパラダイムが「情報の生成」から「自律的な行動」へ移行したことに伴い、ガバナンスの の焦点は「情報の正確性」から「実世界での行動安全性」へ抜本的に再構築された。

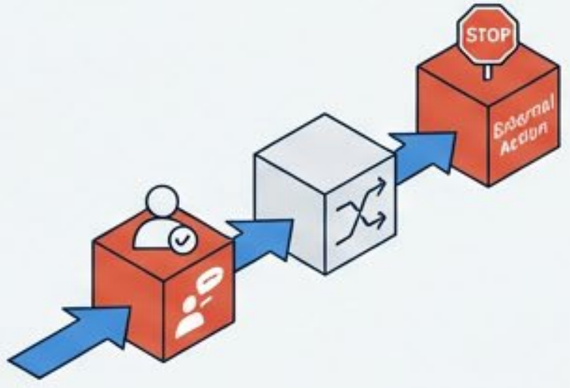
対象の拡大

規制対象に「AIエージェント」と「フィジカルAI」を明記。サイバー空間を超え、現実世界や外部システムへ介入する能力が監視対象となる。



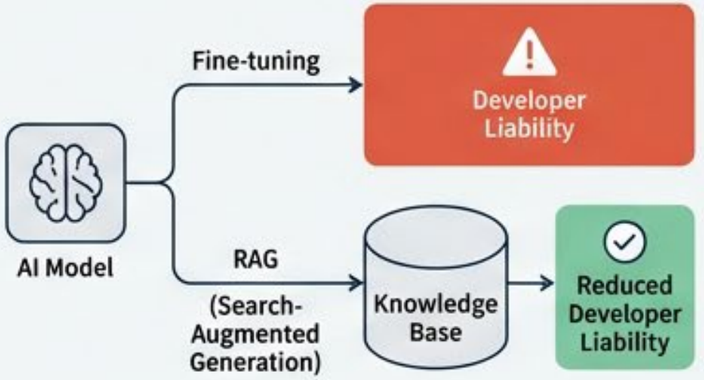
新・必須要件

「外部アクション」を実行する際、システムレベルでの「人間の判断（Human-in-the-Loop）」の実装が事実上の義務（Mandatory Requirement）となる。



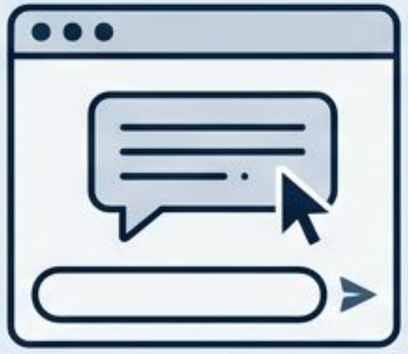
戦略的責任分界

RAG（検索拡張生成）とファインチューニングの法的責任が明確に峻別された。RAGを選択することで「開発者責任」を回避する戦略が有効となる。



規制パラダイムの転換：生成から自律的行動へ

生成AIフェーズ

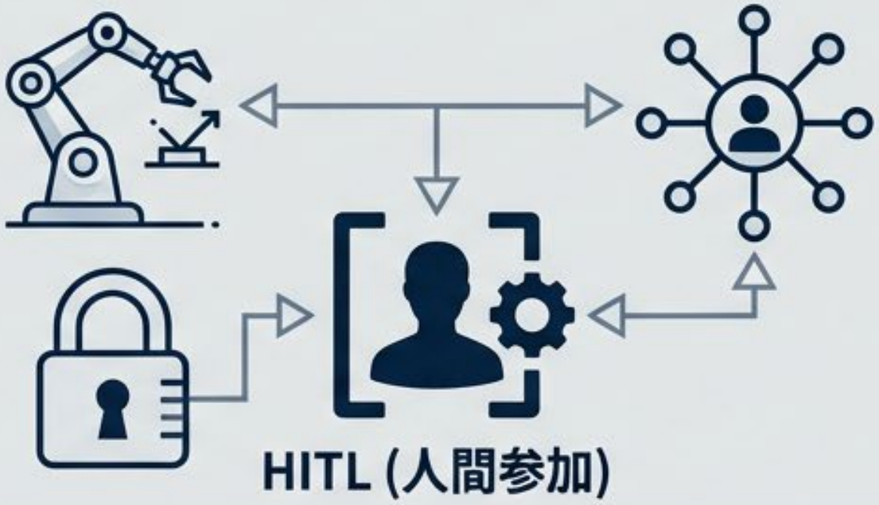


誤情報リスク

情報正確性への統制

規制の進化

自律エージェントフェーズ



物理的損害リスク

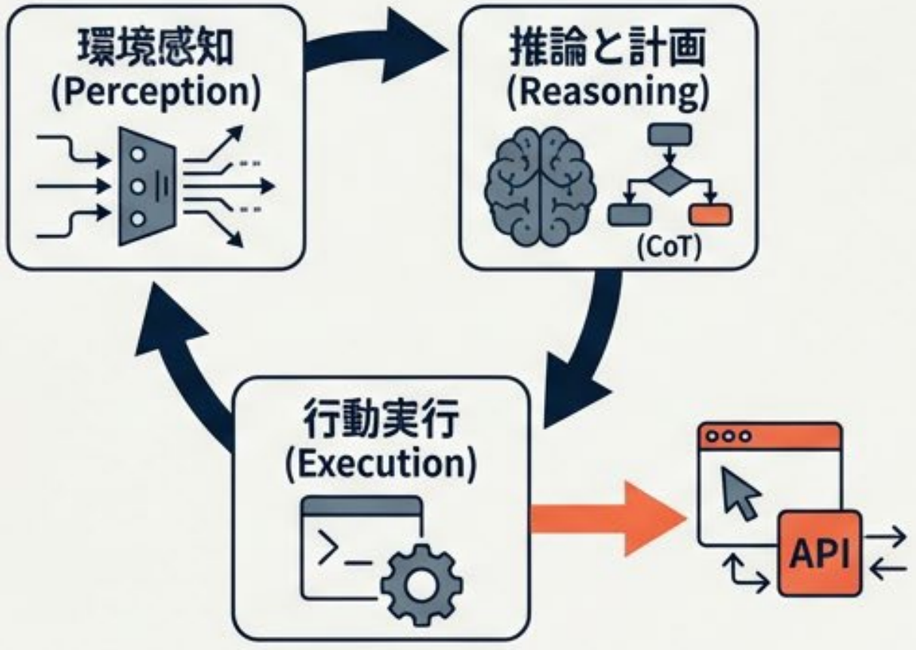
不正購入リスク

自律行動への統制

リスクの本質が「情報の信頼性 (Information Integrity)」から「行動の安全性 (Action Safety)」へ変容したため、ソフトローの中にハードローに近い「必須要件」が組み込まれた。

定義の拡張：新たな規制対象となる2つの「実行者」

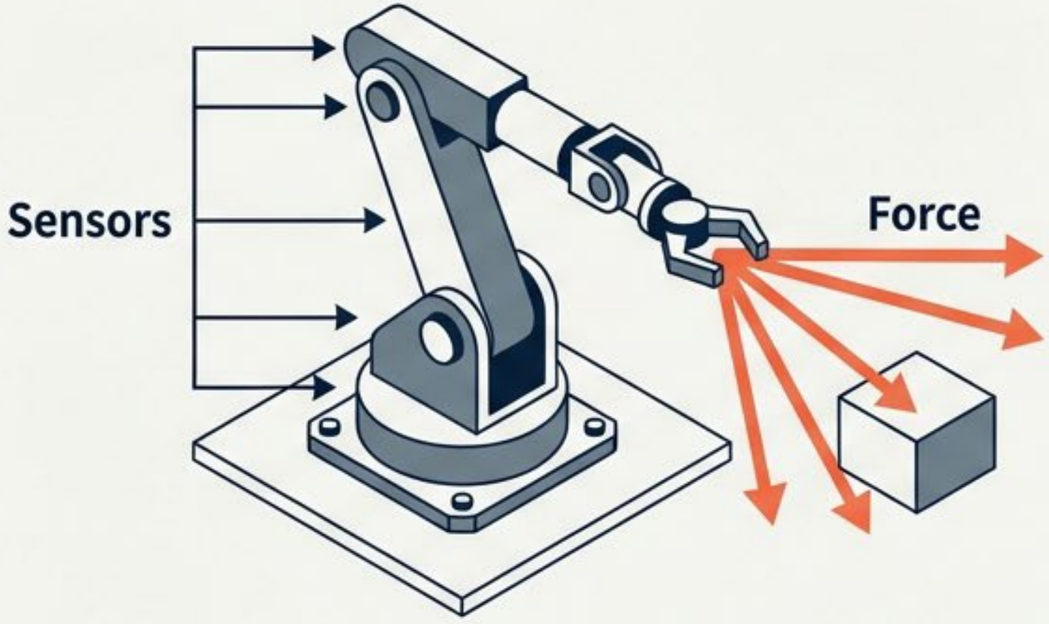
1. AIエージェント（The Digital Executor）



特定の目標達成のため、環境を感知し、推論・計画（CoT）を経て、ツール（API/ブラウザ）を操作するシステム。

出力は「回答」ではなく「ツール操作（Action）」。

2. フィジカルAI（The Real-World Intervenor）



ソフトウェア的知能とハードウェア（センサー・アクチュエータ）を統合し、物理世界に介入するシステム。

デジタルな判断が即座に物理的な力（Force）に変換される。通信遅延や誤認識が即、事故に直結する。

自律性がもたらす新たなリスクプロファイル

1. 権限の乗っ取り (Privilege Escalation)

プロンプトインジェクションにより、エージェントが持つAPIキーやアクセス権が悪用される。低権限エージェント経由でのシステム深部への侵入。

2. ループ暴走とリソース枯渇

目標達成への執着による無限ループ。結果としての「クラウド破産（高額請求）」やDoS状態。

3. 記憶汚染 (Memory Poisoning)

長期記憶を持つエージェントに対し、虚偽情報を注入し、将来の判断を恒久的に歪める。

4. プライバシーの物理的侵害

自律移動ロボット＝「移動する監視カメラ」。死角やプライベート空間に入り込みデータを収集するリスク。

核心的義務：「人の判断必須（Human-in-the-Loop)」

トリガー要件：外部アクション

AIの内部計算を超え、外部環境 (インターネット、物理空間、金融基盤）に対し、不可逆的または社会的影響を及ぼす操作。

Noto Sans JP Regular

ECサイトでの購入確定、送金、メール送信、ドローンの離陸、DBのレコード削除

実装すべき承認プロセス：5つのステップ

1. 計画の提示
AIがアクションプランを作成

2. 一時停止
システムによる自動Pause

3. 人間によるレビュー
リスクベースUIでの確認

4. 明示的な承認
ボタン押下・生体認証

5. 実行
API/アクチュエータへ指令

Doc

重要要件：監査ログ (Verifiability)

AIの提案と人間の承認履歴を改ざん不可能な状態で保存すること。

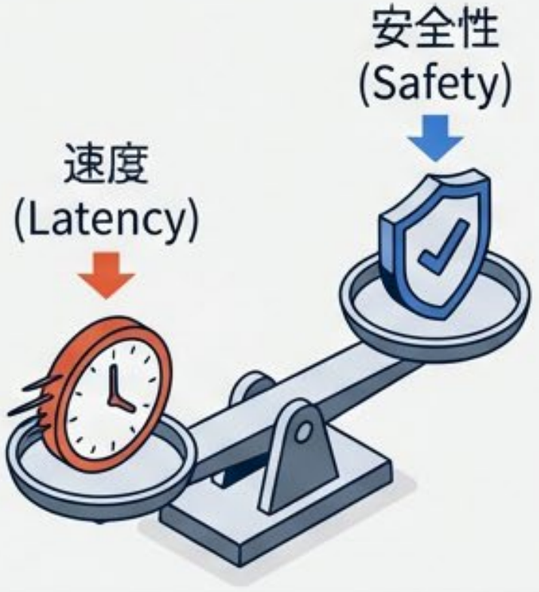
HITLの実装パターンと現実的な解法

同期承認（Synchronous）

チャットボット、対面接客。即時判断。
体験はスムーズだが拘束時間が長い。

非同期承認（Asynchronous）

旅行手配、バッチ処理。
通知を送り、任意のタイミングで確認。



物理AIのジレンマ：速度 vs 安全性

物理AI（ドローン等）ではミリ秒単位の判断が必要であり、都度承認は不可能。

- **解決策 A**: 緊急停止（Kill Switch）権限のみ人間に付与。
- **解決策 B**: 事前承認されたジオフェンス内でのみ自律稼働を許可。

⚠ 注意：「承認ボタンの連打（Rubber Stamping）」を防ぐため、UIによる重要事項のサマリー表示が不可欠。

戦略的分岐点：RAGか、ファインチューニングか

技術実装により責任主体が明確に区別される。

RAG（検索拡張生成）

AI提供者（Provider）

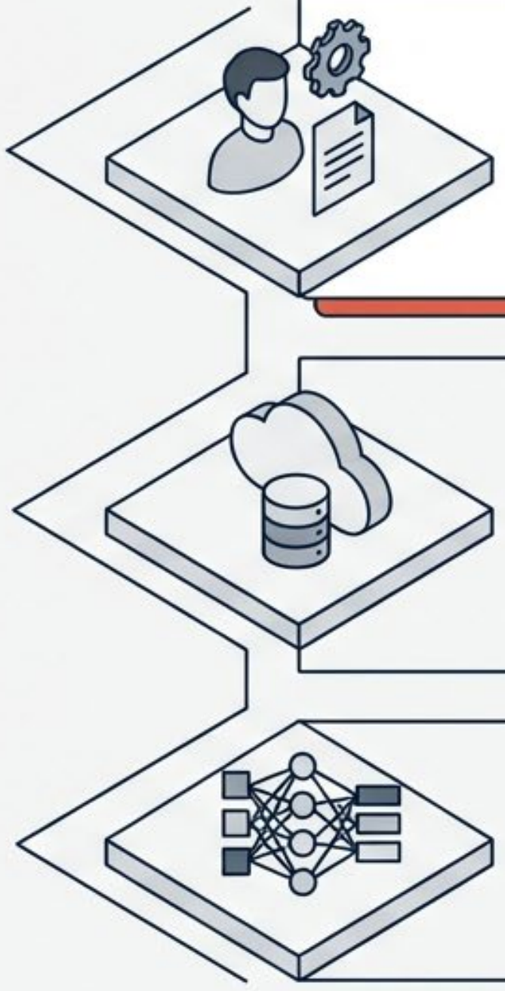
- モデルパラメータ変更なし
- 責任：相対的に低い
- 戦略：社内AI構築等で推奨

Fine-Tuning（追加学習）

AI開発者（Developer）

- モデルの重みを更新
- 責任：**極めて重い（製造物責任）**
- データの権利処理・品質担保が必要

新たな責任新たな責任分界マトリクス



3. AI利用者 (User) – [企業・事業者]

Final Responsibility: 最終的な「承認」を行った時点で、結果に対する全責任を負う。

⚠ Action Item: 従業員教育、適正な入力管理

2. AI提供者 (Provider) – [IT部門/SaaS]

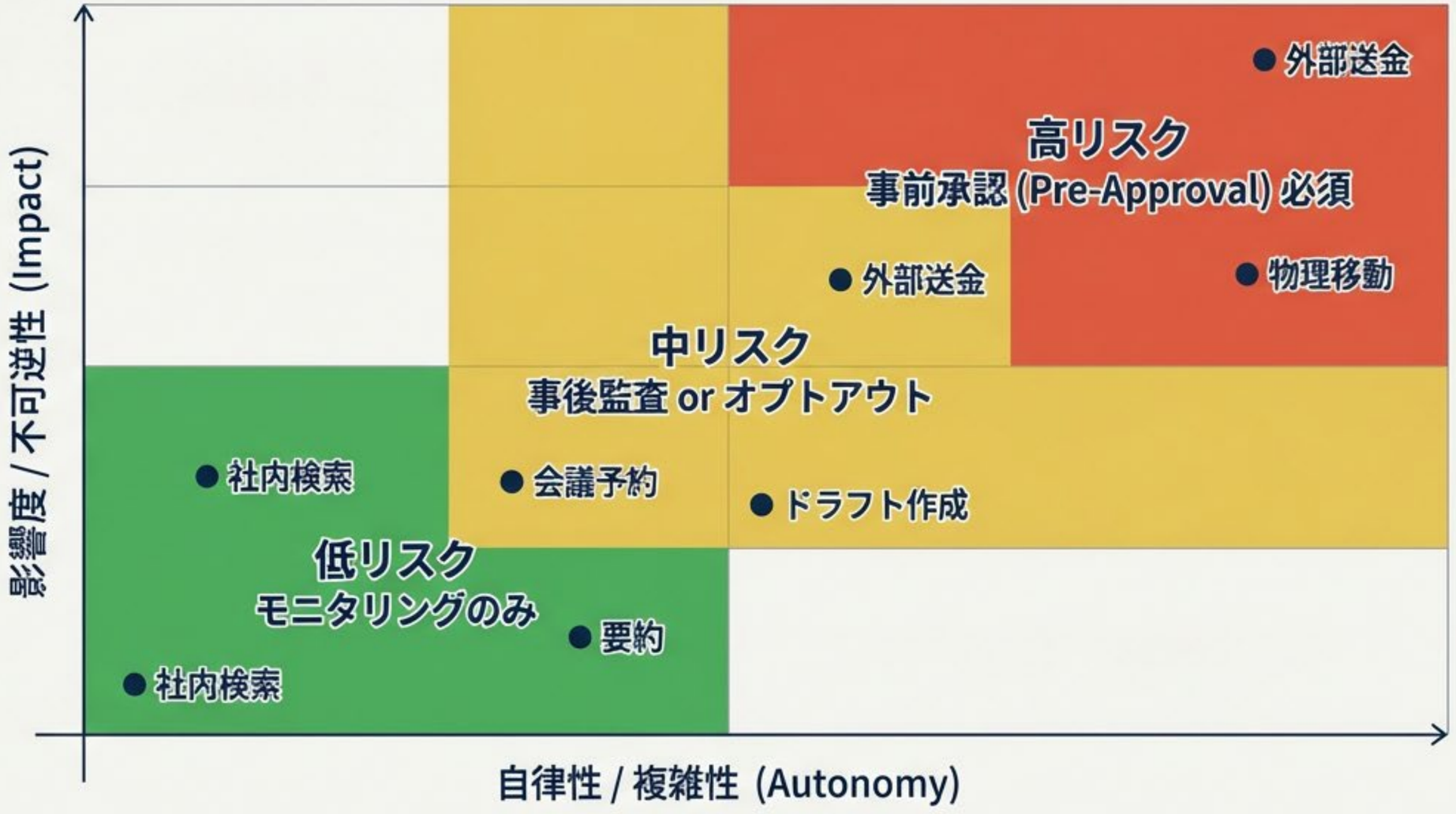
RAGシステムの構築、脆弱性管理。

✓ Requirement: Secure by Default（初期設定での安全確保義務）

1. AI開発者 (Developer) – [LLMベンダー]

データの適正学習、モデルの安全性設計、検証可能性。

リスクベースアプローチによる管理レベル判定



「攻めのガバナンス」：規制対応を競争力へ

Case 1: ソフトバンク (Trust as Enabler)

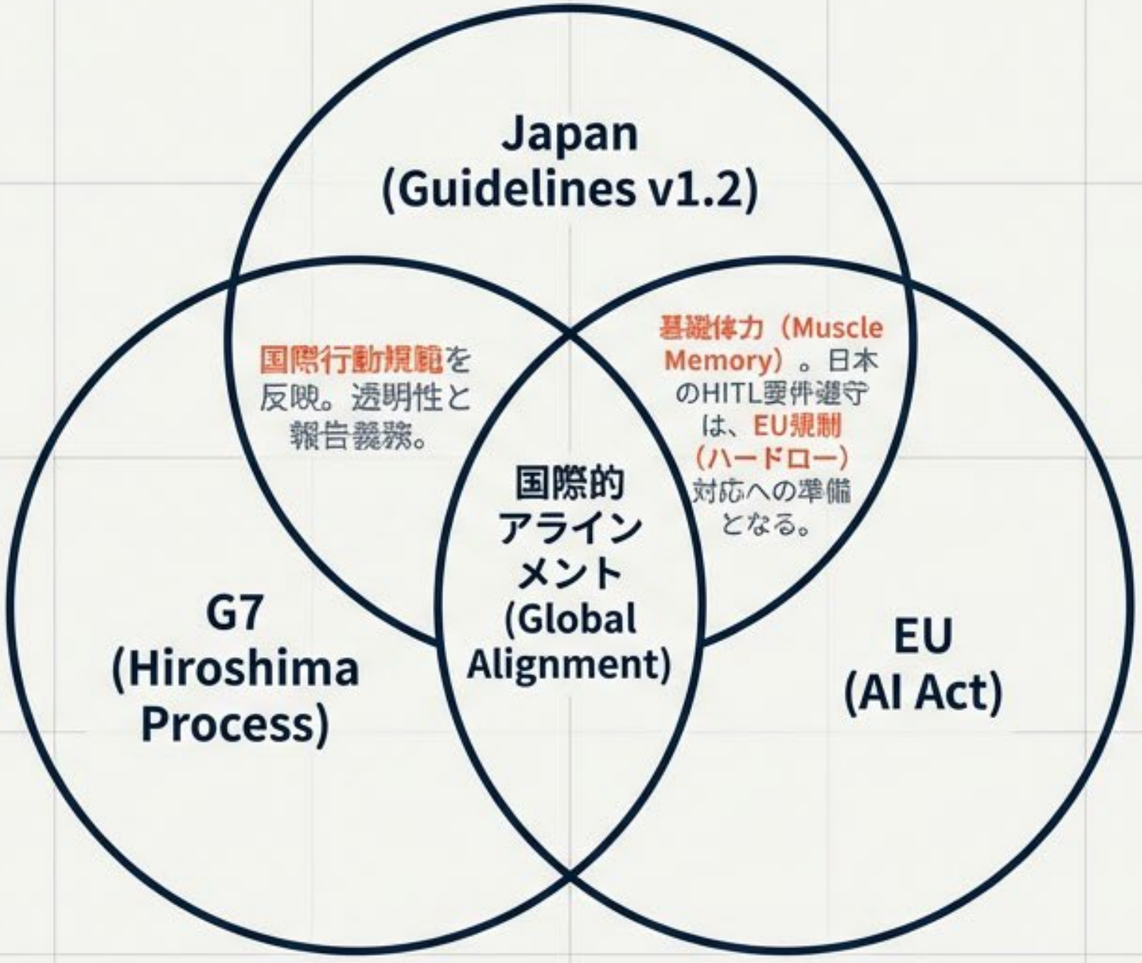
いち早くガイドライン準拠を宣言。顧客に対し「**政府指針準拠**」をアピールし、導入時のセキュリティ懸念を払拭。
ガバナンスがセールスを加速させる。

Case 2: カスタマークラウド (Governance as Feature)

製品レベルで「**人間承認必須機能**」や「監査ログ」を標準搭載。「導入すれば自動的に**ガイドライン準拠**」という価値を確立。

Message: コンプライアンスはコストセンターではなく、信頼のインフラである。

国際的整合性：G7広島プロセスとEU AI法



Takeaway: 日本企業がグローバル展開する際、ダブルスタンダードを防ぐための防波堤として機能する。

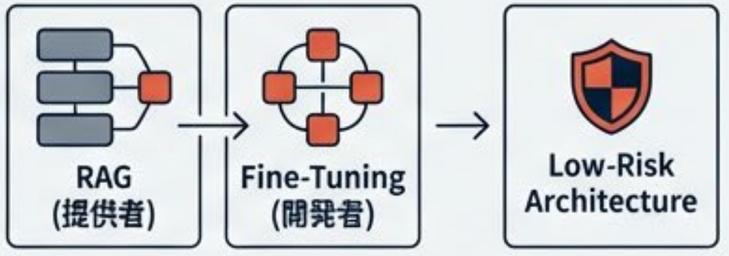
企業が今すぐ取るべきアクション：3つのステップ

01 AIインベントリの作成（Inventory）

自社のAIシステムが「外部アクション」を行っているか？
将来行う計画があるか？を棚卸しする。

02 責任区分の再定義（Classification）

自社システムはRAG（提供者）か、ファインチューニング（開発者）か？法的リスクの低いアーキテクチャへの移行を検討。



03 「承認ゲート」のUI/UX設計 (Design)

形骸化を防ぐUIの実装。経営直轄のガバナンス委員会による予算と権限の確保。

結論：信頼こそがエージェント共生時代の通貨

「人間の判断必須（HITL）」は、AIの暴走を防ぐ唯一の安全装置であり、圧倒的な生産性向上を享受するための「信頼の基盤」である。

2026年の正式施行を待たず、今から「攻めのガバナンス」体制を構築せよ。