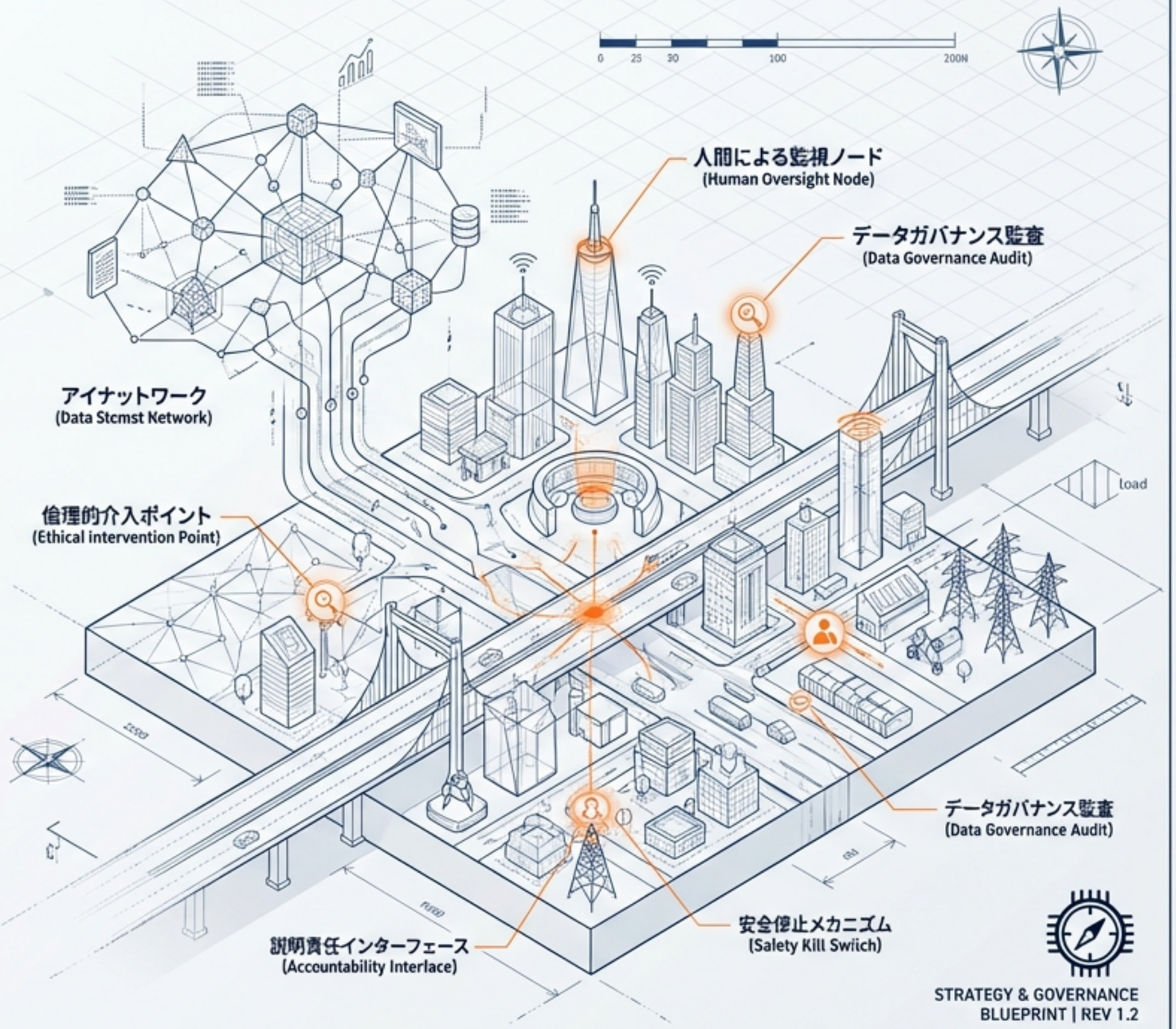
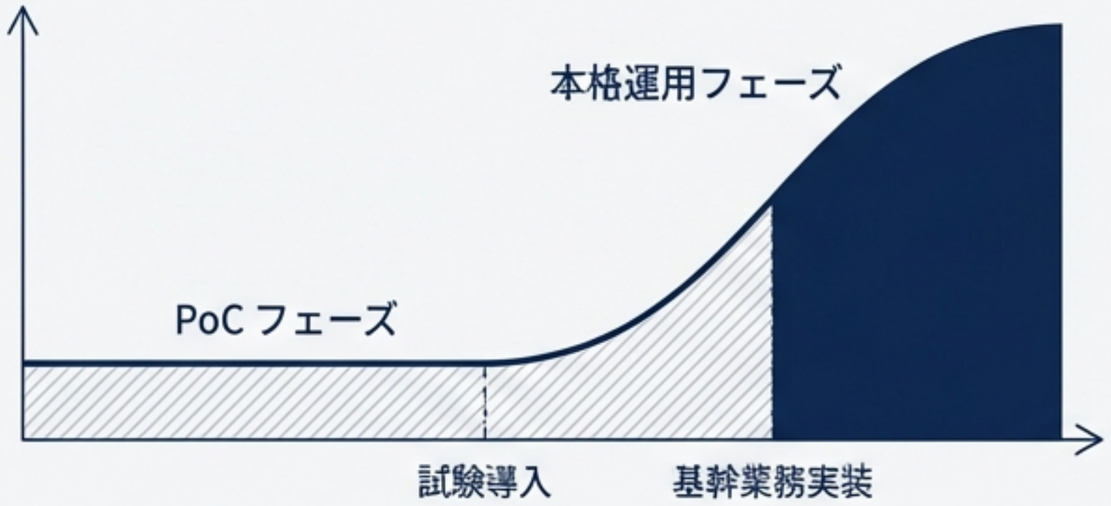
経営層・法務・AIプロジェクトリーダー向け特別ブリーフィング

自律型AI時代のガバナンス再定義

総務省・経済産業省「AI事業者ガイドライン（第1.2版）」の戦略的解読と実践ブループリント



PoCから本格運用フェーズへ

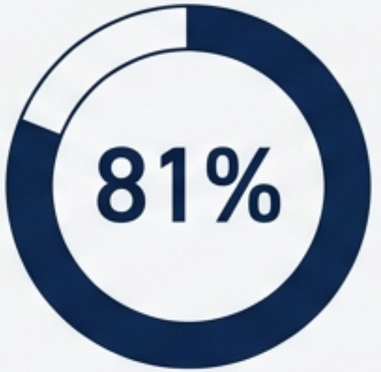


生成AIは試験導入を終え、企業の基幹業務・社会インフラに直接組み込まれる段階へ突入。

Living Documentとしての進化

- 第1.0版 (2024.4)
- 第1.01版 (2024.11)
- 第1.1版 (2025.3)
- 第1.2版 (2026.3)

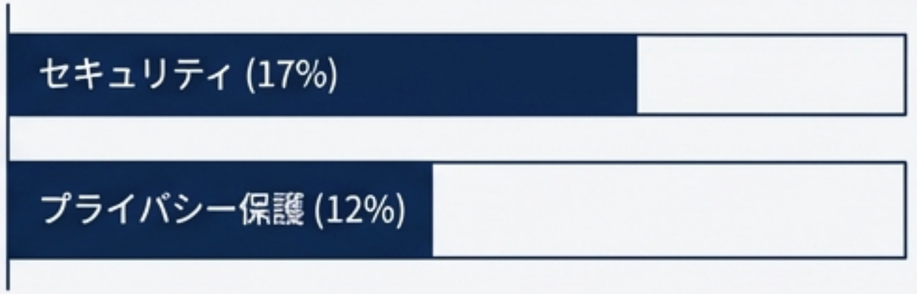
市場の現在地（事前アンケート調査より）



ガイドラインの認知度

全社的な活用

トップリスクの懸念

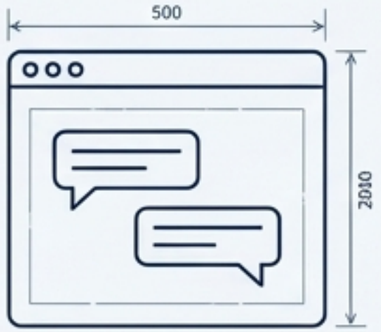


企業は未知の脅威に対する不安を抱えたまま、社会実装のフェーズへと突入している。

「AI事業者ガイドライン（第1.2版）」におけるガバナンス対象の拡張とリスク構造の変容

第1.1版まで

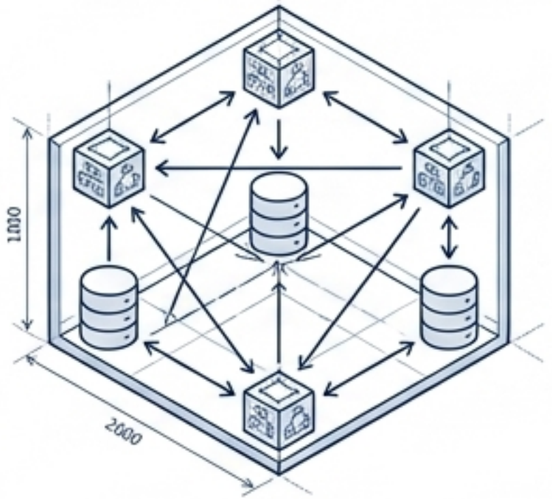
Web上の生成AI
（受動的AI）



リスク：誤情報（ハルシネーション）

第1.2版の追加

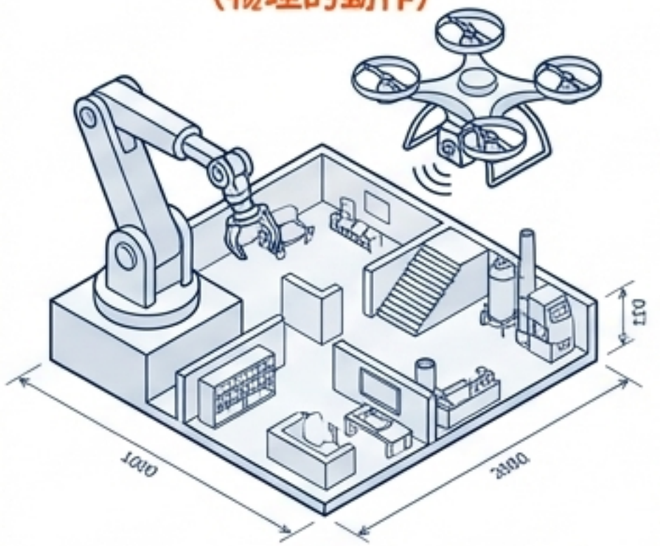
AIエージェント
（自律実行）



リスク：不正行為・誤った意思決定

第1.2版の追加

フィジカルAI
（物理的動作）



リスク：身体的危害（Safety Risk）

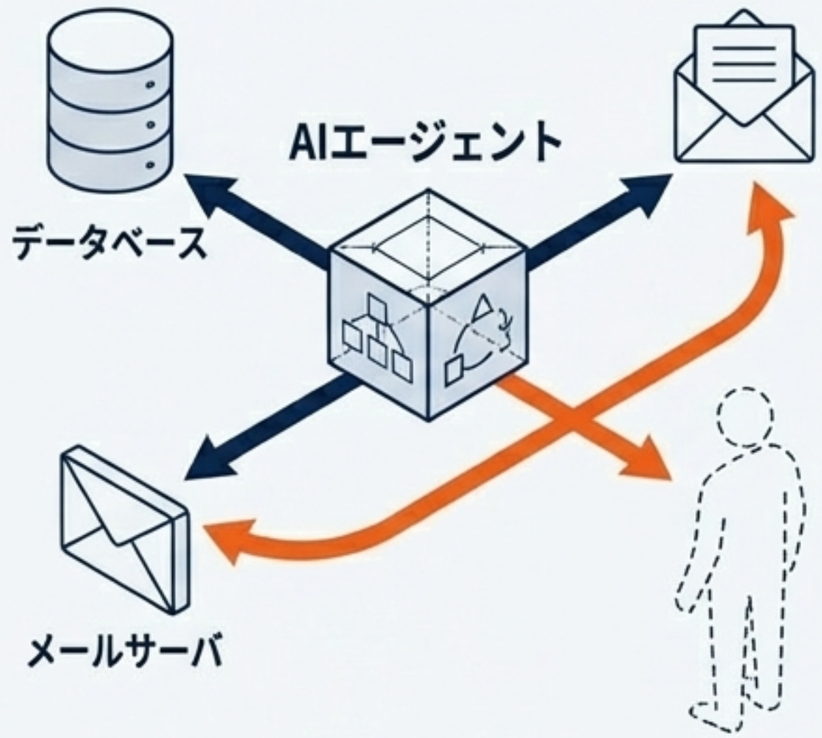
Key Insight

対象が「デジタル空間の受動的ツール」から「物理・社会へ直接作用する自律システム」へ劇的に拡張。
これに伴い「人間の関与（HitL）」が絶対的な要請となった。

企業の基幹業務・社会インフラへの直接組み込みが進み、未知の脅威に対する不安を抱えたまま、社会実装フェーズへと突入している。

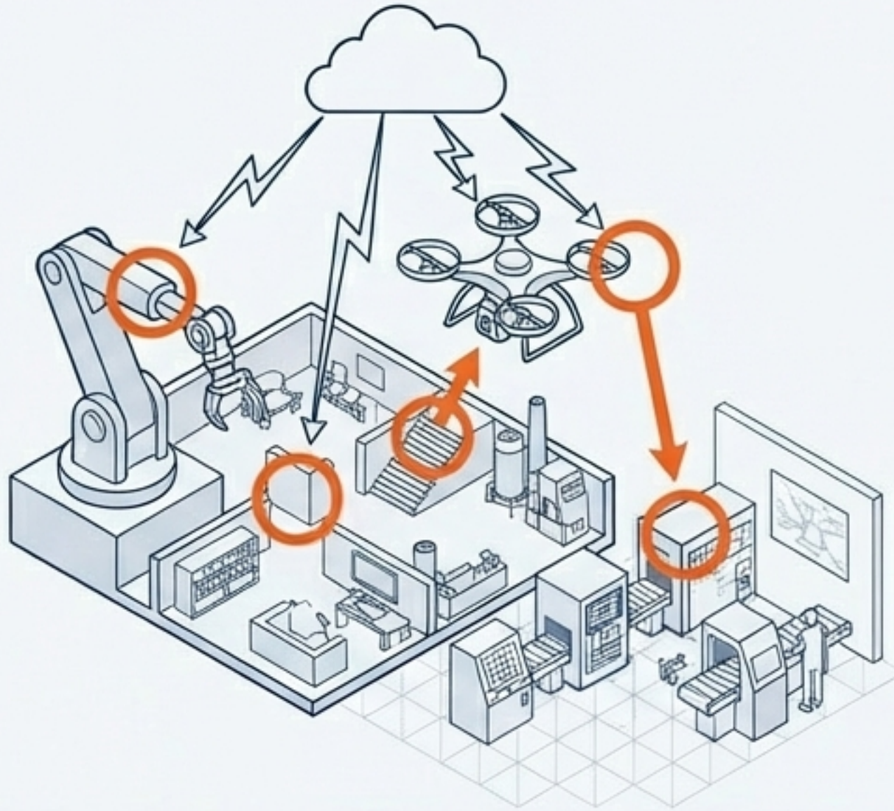
新たな脅威空間へのディープダイブ

AIエージェントの暴走リスク



ツール（社内DB、メール、API）の自律的操作による未曽有のリスク。
例：誤った見積書の自動送信による経済的損失。
アルゴリズムのバイアスによる特定候補者の自動不採用（法的責任・信用失墜）。

フィジカルAIの安全性再定義



センサー・アクチュエーターを介した物理空間への直接干渉。
サイバー攻撃が直接的な「物理的危害」や「人命への脅威」に転化。
従来のソフトウェア開発の枠を超えたハードウェア連携の安全性評価が必須。

システムアーキテクチャの核心：Human-in-the-Loopの義務化

アラート・オレンジの関所

自律的処理ゾーン
（効率化の最大化）

データの分析、ドラフト作成

人間の最終承認
(Human-in-the-Loop)

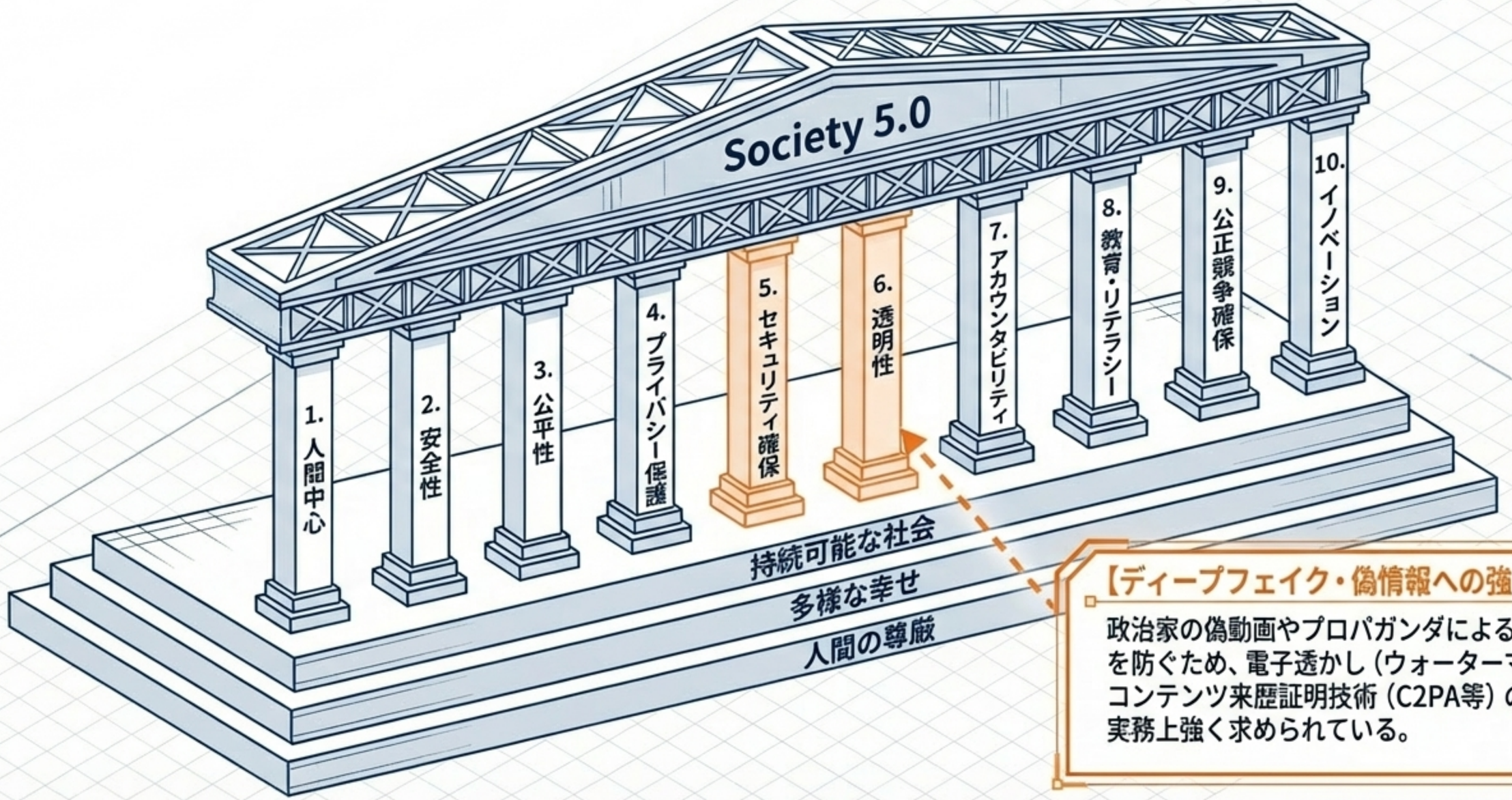
不可逆的アクションゾーン

外部へのメール送信、
基幹DBの更新、契約の締結

アカウンタビリティの担保

「オートメーション・バイアス」の排除。AIは法的責任を負えない。
人間を最終承認者として配置することで、インシデント発生時の「AIが勝手にやった」という責任逃れを不可能にし、法人としての説明責任を担保する。

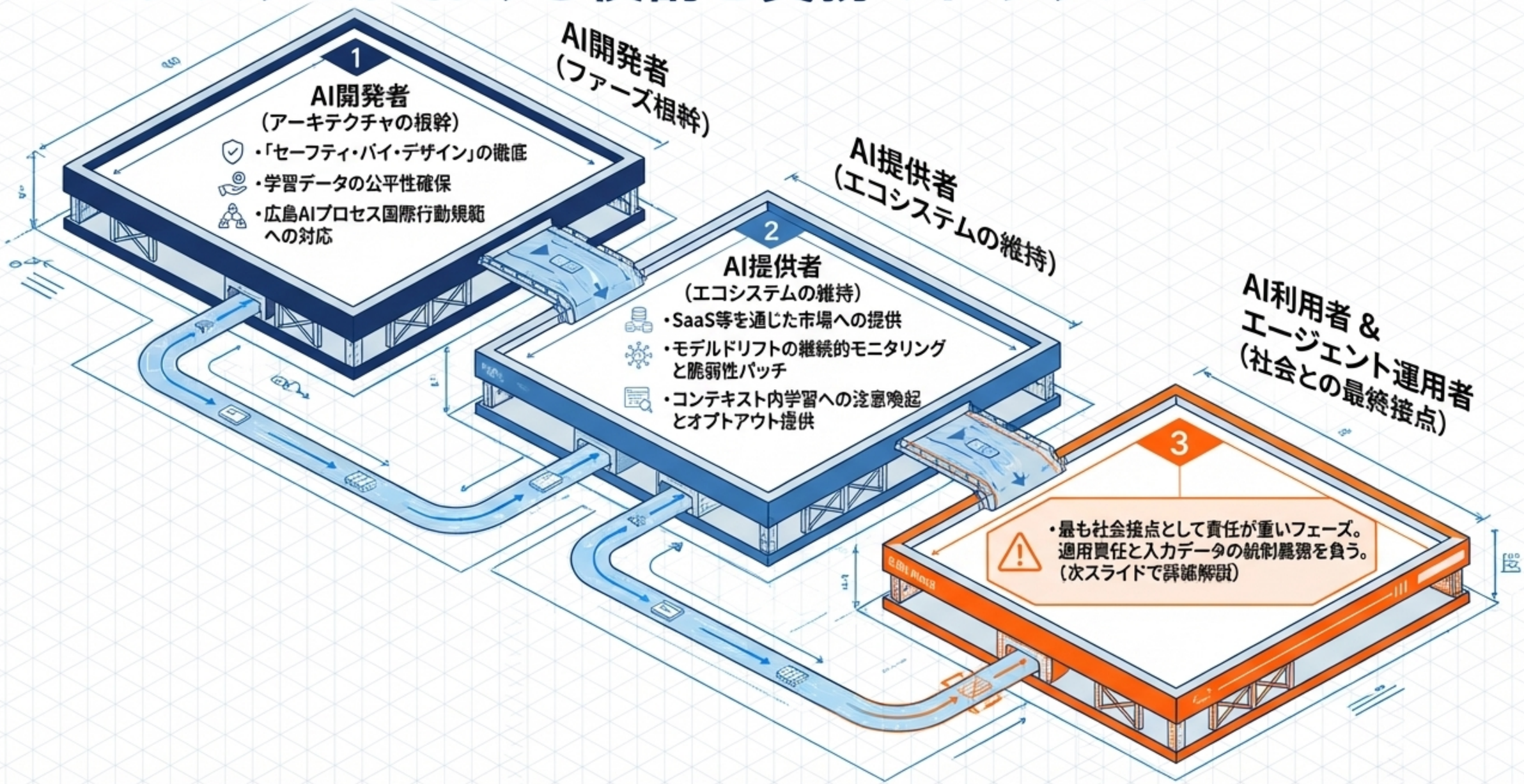
ガバナンスの哲学的基盤と10の共通指針



【ディープフェイク・偽情報への強い警戒】

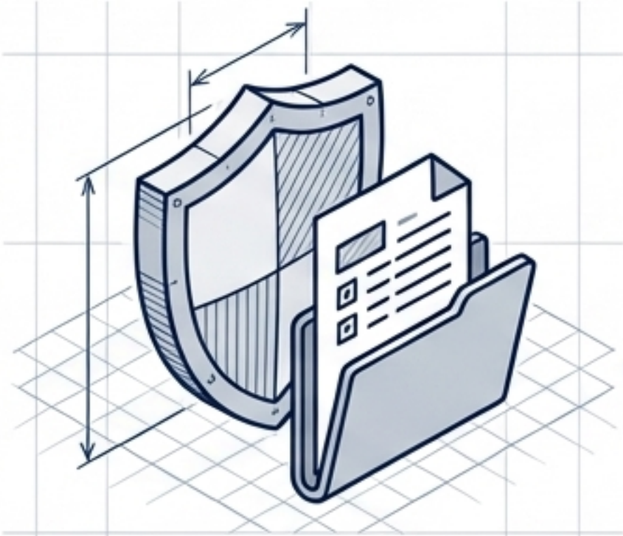
政治家の偽動画やプロパガンダによる社会混乱を防ぐため、電子透かし（ウォーターマーク）やコンテンツ来歴証明技術（C2PA等）の導入が実務上強く求められている。

エコシステムにおける役割と責務マトリクス



利用者の重い適用責任：入力から出力までの厳格な統制

入力データの厳格統制



- 従業員によるプロンプト入力の監視。
- 顧客の個人情報、未公開財務情報、営業秘密の入力防止。
- 強固なデータガバナンスとリテラシー教育の徹底。

AIモデル（プロセッシング）

AIモデル（プロセッシング）

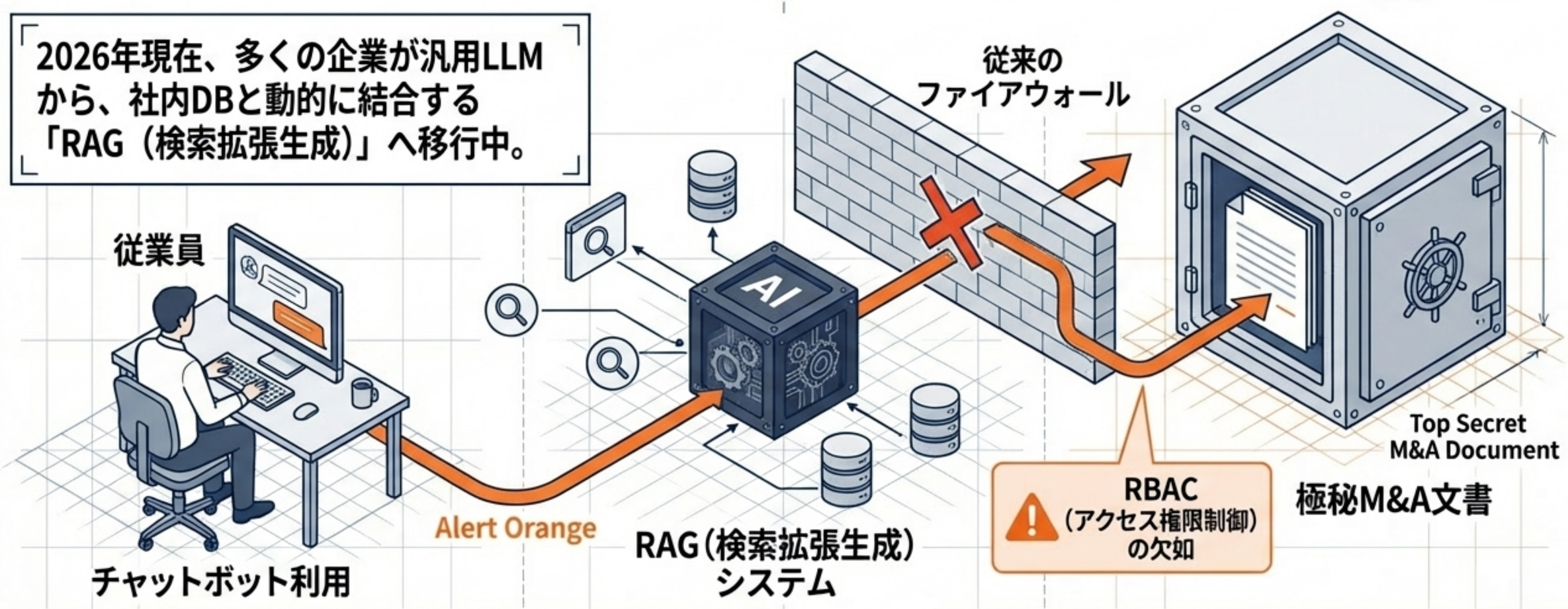
出力結果の管理と適用責任

- 盲信の排除（ハルシネーション・バイアスの検証）。
- 自社の事業判断としての「合理的な最終判断（HitL）」。
- ステークホルダーに対する「専用の問い合わせ窓口」の設置。

Key Insight: 自ら開発を行わなくても、最終的な意思決定として社会・顧客に接するため、利用者の適用責任は極めて重い。

迫り来る新たなサイバー脅威：RAGパラドックス

2026年現在、多くの企業が汎用LLMから、社内DBと動的に結合する「RAG（検索拡張生成）」へ移行中。



RAGの無秩序な利用により、外部からの攻撃よりも、内部システムでの「AIを介した権限のすり抜け」が最大の脅威となっている。

SIerやベンダーとの契約において、「どのレベルの機密データが、どの段階で処理されるか」の厳密な切り分けが急務。

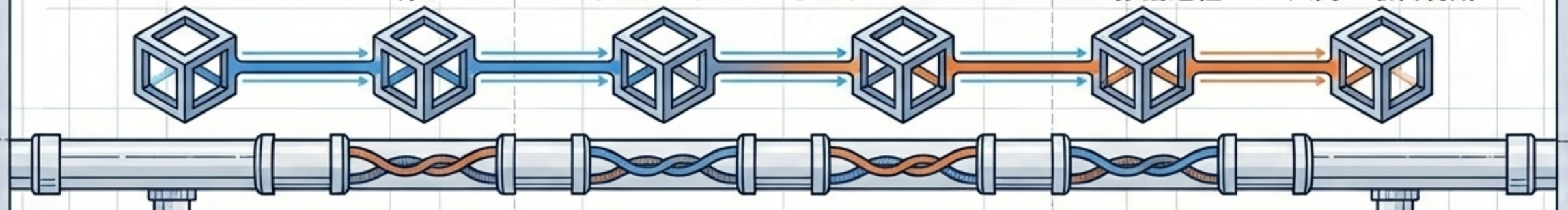
技術的解決策：データエンジニアリングと完全なトレーサビリティ

現状のサイロ化・属人化

ログの分断により監査不適合

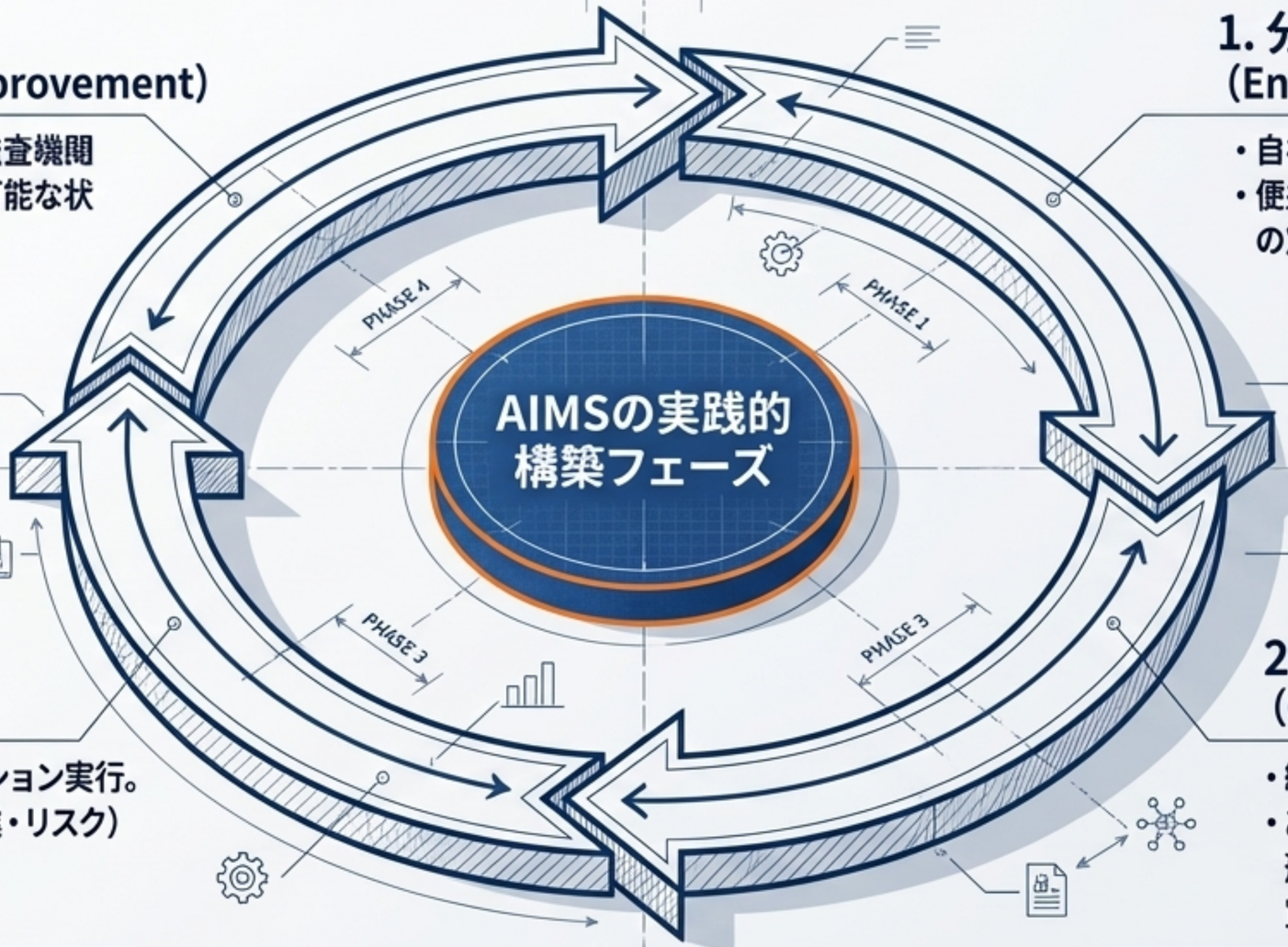
データ統合基盤による完全な追跡可能性

いつ　誰が　どのデータで　どんなプロンプトで　AIの推論過程　人間の最終判断



インシデント発生時に意思決定チェーンを完全に事後再構築できる、
高度なデータ統合基盤（データカタログ＋ETL）体制が必須。

AIMS（AIマネジメントシステム）実装ロードマップ



1. 分析
(Environment & Risk)

- 自社のAI習熟度の把握。
- 便益とリスク（社会受容性）の定量的・定性的評価。

2. ゴール設定
(Goal Setting)

- 経営陣のコミットメント。
- あるべきAIガバナンス最終形態の定義と社内外への宣言。

3. デザイン
(System Design)

- ギャップを埋めるアクション実行。
- 部門間（法務・IS・事業・リスク）のサイロ破壊。
- 全社リテラシー教育。

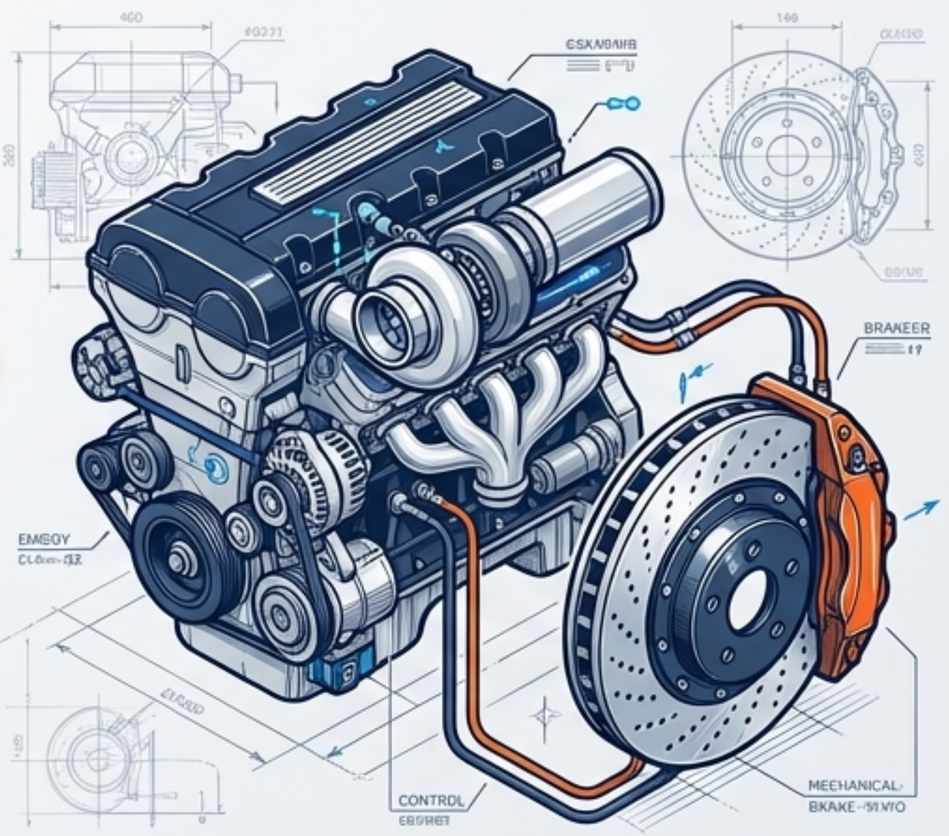
4. 運用
(Operation & Improvement)

- ステークホルダーや監査機関に対する「常に説明可能な状態」の維持。

マインドセットの転換：ブレーキからイノベーションのアクセルへ

過去のパラダイム：ブレーキ（管理・統制）

第1.2版のパラダイム：アクセル（競争優位の源泉）



- マインドセット: ガバナンスはイノベーションを阻害するコスト。
- リスク: 無法地帯でのシステム開発は、後から規制や反発で運用禁止になる「最大のサンクコスト（埋没費用）リスク」を孕む。

- マインドセット: 明確なガードレールによる事業上の「予見可能性」と「社会的信頼」。
- 結果: 強力で信頼性の高いブレーキがあるからこそ、未知の技術的挑戦に対しても安心してアクセルを全開に踏み込める。

中小企業を含む全事業者が今すぐ取るべき3つの戦略的アクション

1

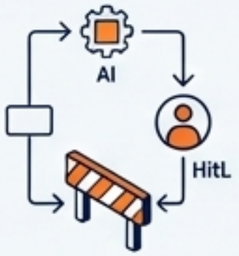
AI利用状況の網羅的な棚卸し

シャドーAI（会社非公認ツールの独自利用）の特定と排除。社内でででの利用状況を完全に把握する全社インベントリ（台帳）の作成。

2

責任範囲と「自律の境界線」の明確化

業務プロセスを細分化。「AI単独で完結させてよい業務（要約等）」と「必ず人間の承認（HiL）を挟むべき業務（外部送信・データ更新等）」を厳密に線引き。



3

社内ガイドラインの策定と継続的更新

ツールのホワイトリスト、機密レベルに応じた入力禁止情報の定義、問題発生時（ハルシネーション・漏洩疑疑い）の緊急対応フローの文書化。

機械の圧倒的効率性と人間の倫理的責任の調和

AI事業者ガイドライン（第1.2版）は、単なる法務部門のルールブックではない。自社のデータ基盤を根本から整備し、組織の意思決定プロセスを次世代型に再構築するための「戦略的ロードマップ」である。

イノベーションの「アクセル」とリスクマネジメントの「ブレーキ」を有機的に連携させることでのみ、持続可能な競争力と「Society 5.0」の実装が達成される。アジャイルにアップデートし続ける組織だけが、本格的なAI駆動社会を生き残る。