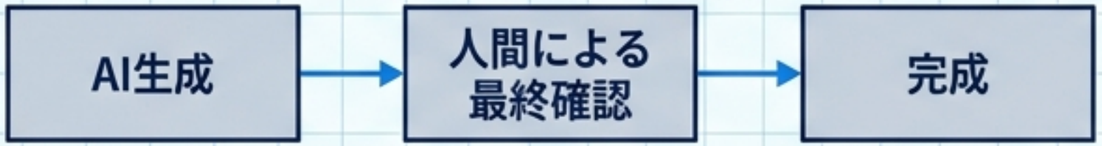
知財AIエージェントの最適設計

「Can / Trust / Own」モデルに基づく権限委譲とガバナンス

Confidential / Internal Strategy Playbook

「最終確認の罠」：既存のHuman-in-the-loopはなぜ破綻するのか



抽象的な運用論による過信

認知負荷の限界

数十億パラメータのLLMが生成した論理的飛躍や事実誤認を、人間が事後的にすべて見抜くことは非現実的。

オートメーション・バイアス

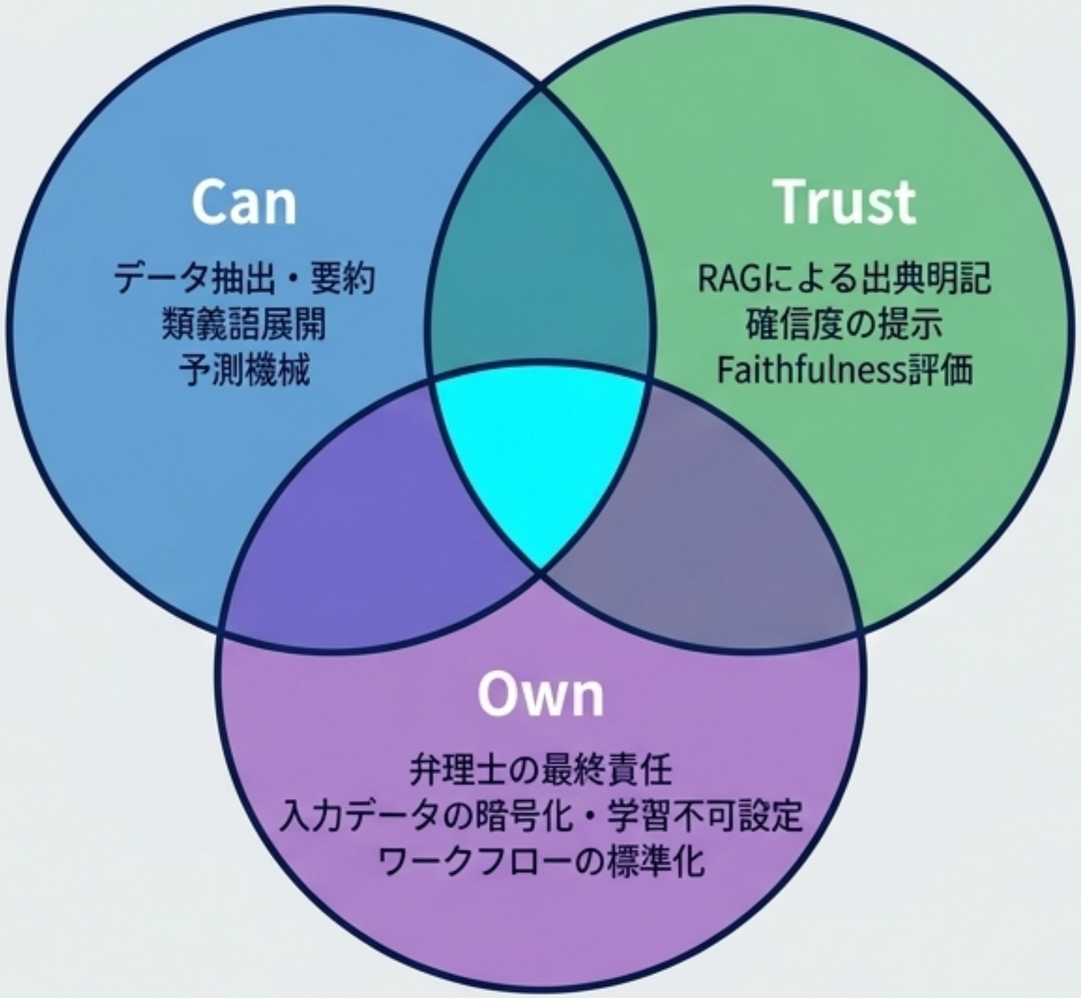
AIの流暢な出力に対する過信による、深刻な権利侵害や特許無効化リスク。

機密情報の漏洩

不適切なプロンプト入力による未公開発明の新規性喪失（致命的インシデント）。

知財業務において、AIに高度な判断を委ねることや、事後的な体裁整備に依存する運用は、ガバナンスとして成立しない。

パラダイムシフト：「Can / Trust / Own」モデルの導入



[Can] 予測能力

膨大なデータに基づく確率論的推論とパターン抽出

[Trust] 工学的検証

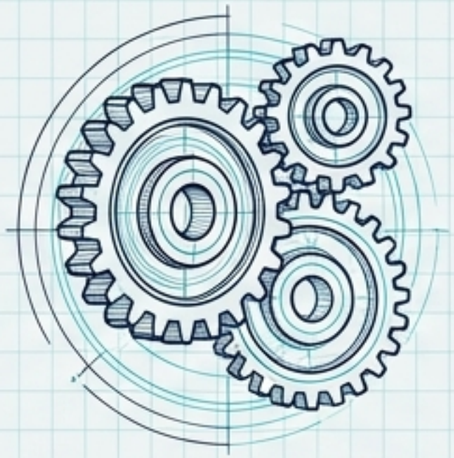
出典明記、確信度提示、Faithfulness評価によるフェイルセーフ

[Own] 責任と統制

弁理士の最終責任、データの暗号化、ワークフローの標準化

知財業務におけるAI活用は、単なる自動化ではなく、予測能力（Can）、工学的検証プロセス（Trust）、および人間による法的責任の保持とワークフロー統制（Own）の三要素が重なり合う領域で設計されなければならない。

[Can] 能力の境界：AIは「思考」しない、「予測」する



AIの領域（確率論的推論）

膨大な学習データに基づき、次に出現する確率の高いトークンを推論。

知財における強み:
文脈的要約（外国語文献からの差異抽出）、パターン抽出（類義語の網羅的展開）。

人間の領域（価値判断と戦略）

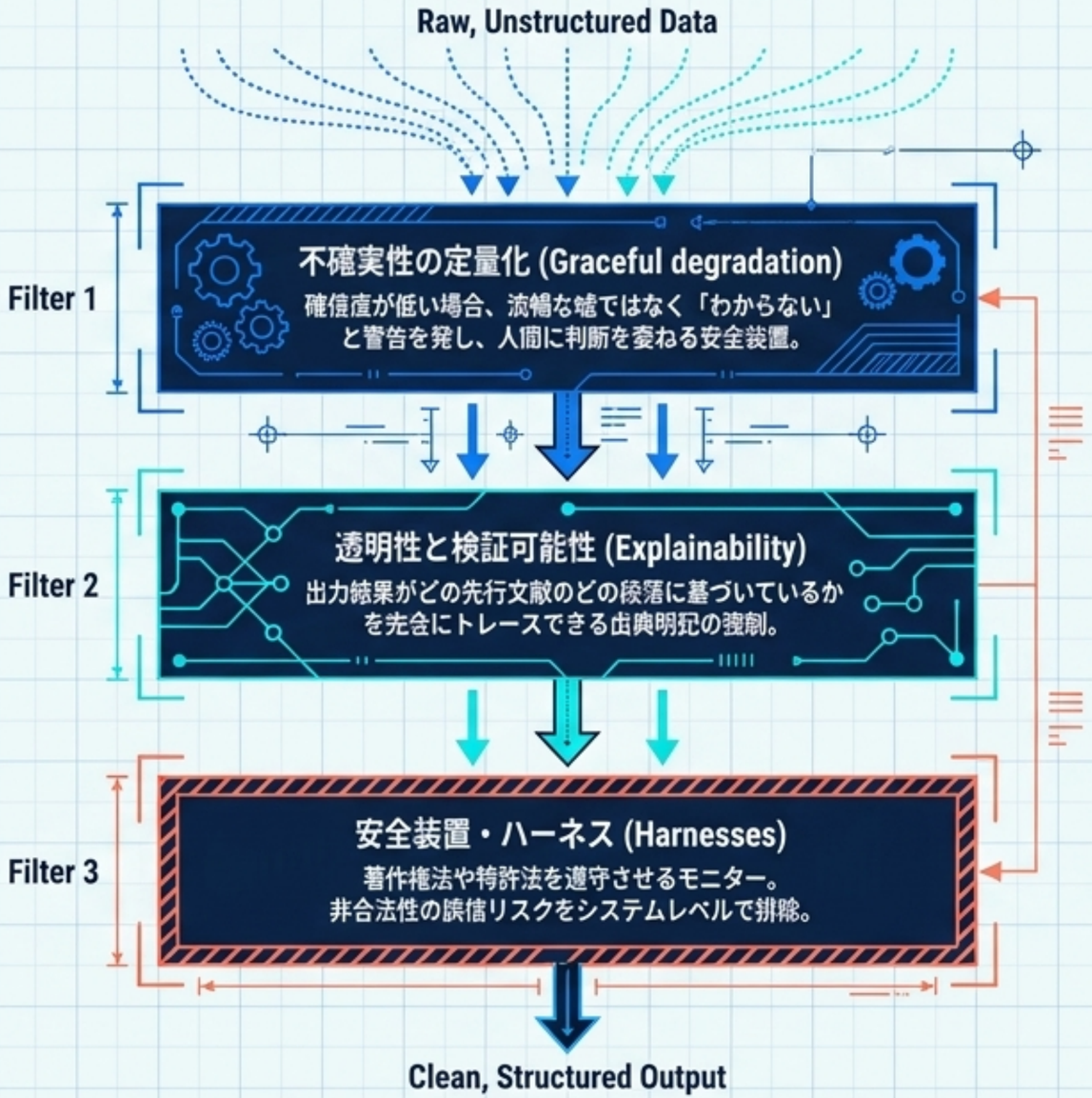
未学習の最新技術（未公開発明）にはハルシネーションの構造的弱点が存在。AIは自律的思考を持たない。

知財における限界:
進歩性の規範的判断、出願国移行の事業戦略決定。

Strategic Action: AIは「完全な解決策」ではない。プロンプトを通じて情報を引き出し（Elicit）、目的関数を整合（アライメント）させるためのコンポーネントである。

[Trust] 信頼の工学：道徳的責任の欠如と「依存」への再定義

AIは道徳的責任を持たない。我々が構築すべきは擬似的なTrustではなく、実績に基づく「Reliance（依存）」のシステムである。



［Own］信託法理に基づく権限委譲とガバナンス

弁理士や知財部門は、企業の無形資産を管理する「受託者（Trustee）」である（日本弁理士会ガイドライン準拠）。

委譲可能な管理機能 (Delegate)

目的: 業務効率化と人間のリソース解放

内容:
- 先行技術の一次スクリーニング
- 関連文書の翻訳
- 明細書の初期ドラフト作成

委譲不可な裁量的義務 (Retain)

理由: 受託者としての根源的な義務。AIへの白紙委任は法格言にも反する義務違反。

内容:
- 進歩性・非自明性の最終判断
- 権利化の可否
- クレームの戦略的解釈

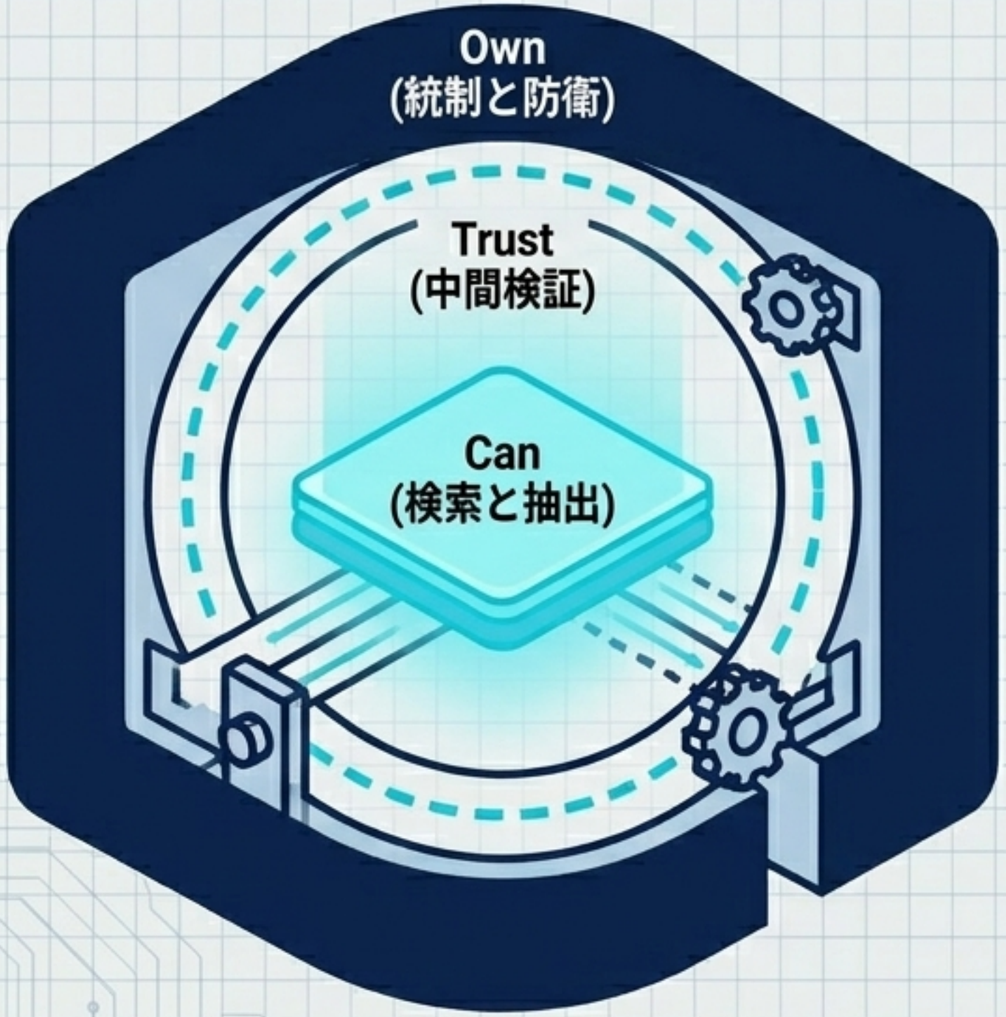
【Owning the Process】

属人的なプロンプト入力を脱却し、自社独自の判断基準を「スキル」としてAIにエンコード（標準化）し、組織として所有せよ。

統合マトリックス：「Can / Trust / Own」の知財実務適用構造

[概念]	[AIに対する解釈]	[ガバナンス要求事項]	[知財実務の具体例]
Can	確率論に基づく高度な予測機械	認識的限界の把握と意図のアライメント	膨大な特許公報からの類義語抽出、定型ドラフト生成
Trust	道徳的責任を伴わないシステムへの依存	不確実性の提示と出力根拠の検証可能性の確保	出典の明記、確信度が低い場合の「わかりません」応答の強制
Own	法的責任の帰属とワークフローの所有	裁量的判断権の保持と自社AIスキルの標準化	AI生成物への弁理士の最終責任担保、未公開発明を保護する専用AI構築

ユースケース1：発明発掘と先行技術調査 (High Own, Moderate Trust)



The Threat: 致命的リスク

パブリックAIへの未公開発明の入力による
「情報漏洩」と「新規性喪失（特許取得不可）」

Own: セキュリティの3つの防衛線

1. エンドツーエンドの強固な暗号化処理
2. 事業者によるデータ解析不可（厳格なオプトアウト）
3. モデル再学習への入力データ不利用（利用規約での完全保証）

Process Architecture: Can & Trust

AIのCan: 関連文献の網羅的検索と類義語展開による高速化。

人間のTrust補完: AI特有の「サイレントリスク（検索漏れ）」を防ぐため、中間プロセスで人間による検索式の妥当性確認を強制配置。

Legal Touchpoint: オープンイノベーション時は残存期間1年等のNDA締結・合意管轄指定など法務連携が必須。

ユースケース2：明細書作成とOA対応（Moderate Can, High Trust）

Goal: 法的文書における一語の差異の致命的リスク（ハルシネーション）を技術的に根絶する。

Standardized System Prompting

「情報源不明・推測の場合は明記し、わからない場合は『わかりません』と答えること」を標準実装。

法的境界線（Legal Governance）：発明者適格性とFTO調査

発明者適格性
（USPTO DABUS事件の教訓）

原則：
AI自身は特許の発明者になり得ない（自然人のみ）。

リスク：
AIの出力をそのまま出願すると、本質的寄与が否定され特許無効化。

対策（Own）：
人間がどのように課題設定し、出力を統合・変形して「構想（Conception）」に至ったか、実験ノート等の開発履歴として詳細に立証・記録するプロセス設計。

FTO・クリアランス調査
（重大リスク業務）

原則：
デジタル庁ガイドラインが定める「重大な影響を及ぼす業務（事業停止・巨額賠償リスク）」。

制約：
非侵害の「法的判断」をAIに委譲することは厳禁。

適用領域：
クレーム構成要件と自社製品仕様の対比マトリックス自動生成など「前処理・構造化タスク」に厳格に限定。

組織実装：CAIOと知財部門の戦略的協働

CAIO (Chief AI Officer)

AI利用ルールとユースケース別リスクレベルの全社的統括。

知財部門 (IP Hub)

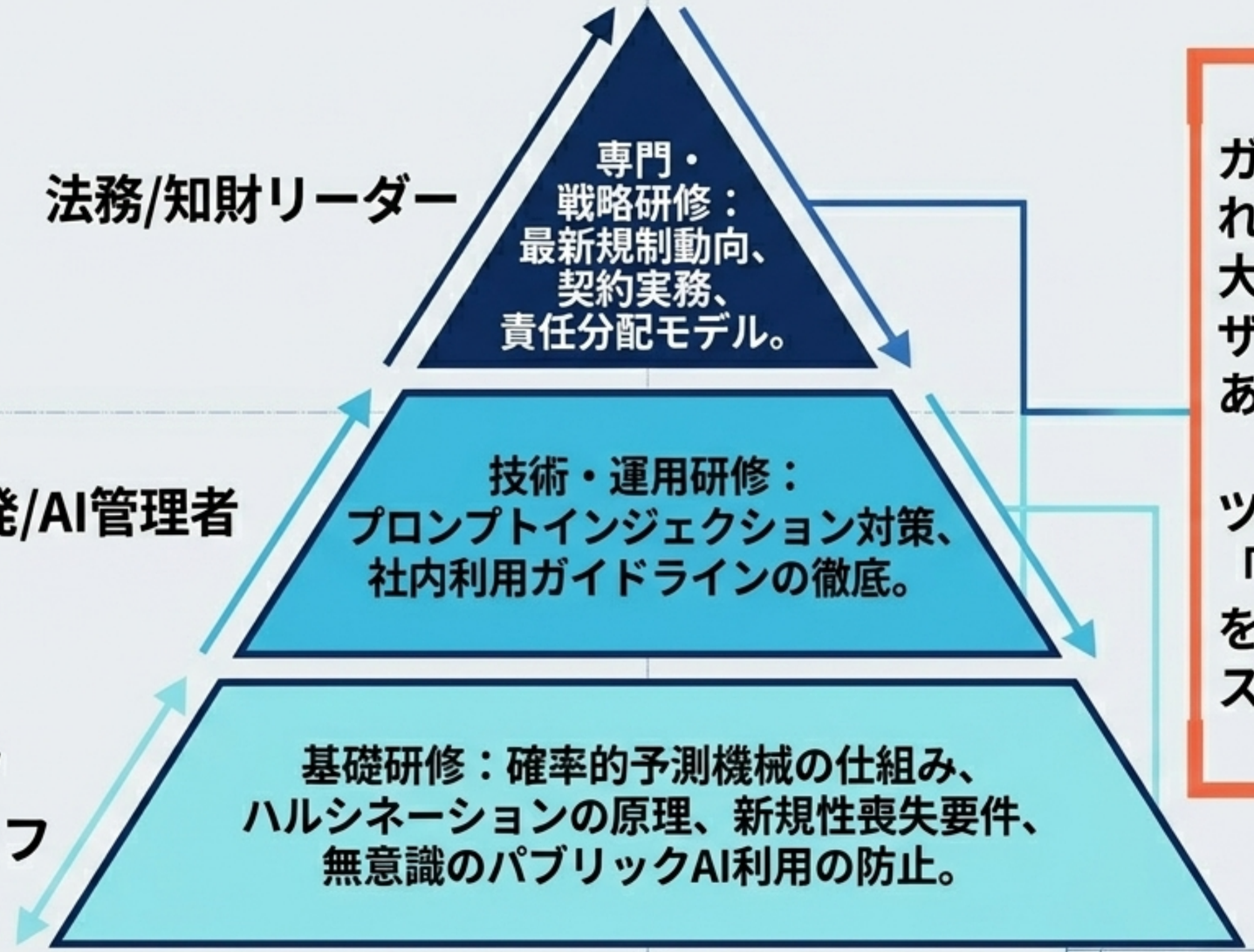
無形資産の保護、著作権/商標/特許の新論点整理、AI生成物の権利リスク管理の主導。

法務部門 (Legal)

企業間契約実務、損害発生時の責任分配の事前調整。

Key Message: 知財部門は単なる「AIユーザー」に留まらず、自社の無形資産保護と第三者権利侵害リスク低減を両立させる全社ポリシー策定の中核（ハブ）を担うべきである。

組織リテラシーの継続的向上：最後の防衛線



ガードレールがどれほど強固でも、最大の脆弱性はユーザーのリテラシーにある。

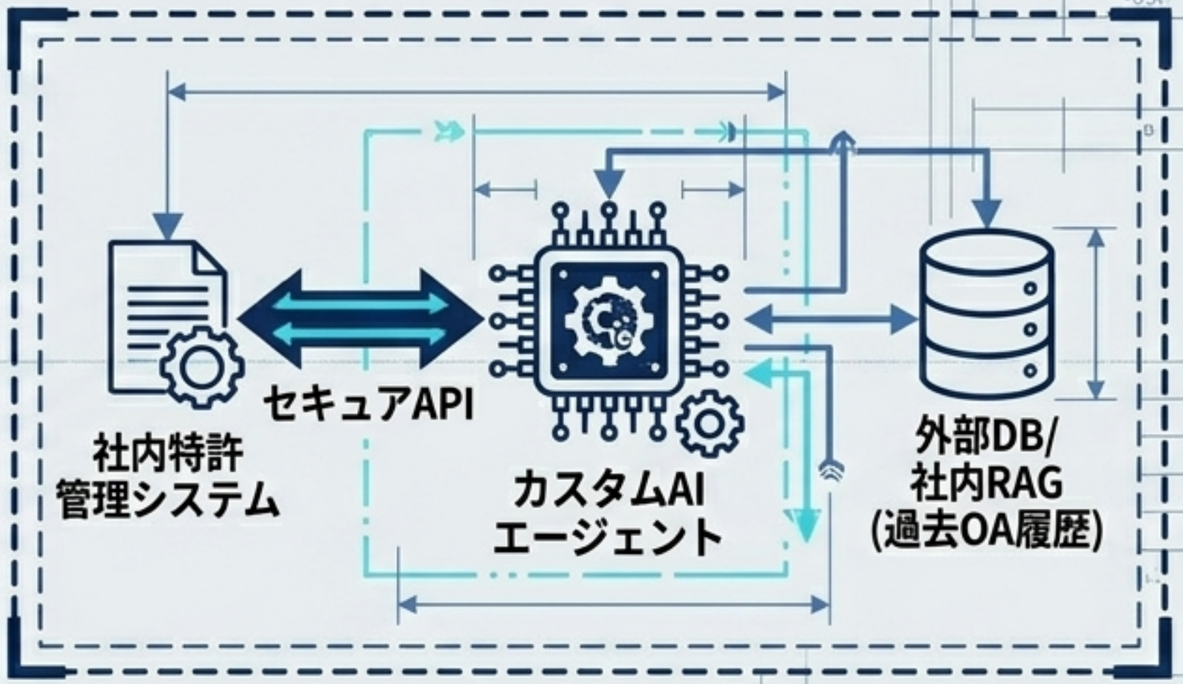
ツール導入前に「予測機械の原理」を組織全体にインストールせよ。

進化：汎用チャットボットから「知財AIエージェント」へ

属人的フロー（Siloed & Inefficient）

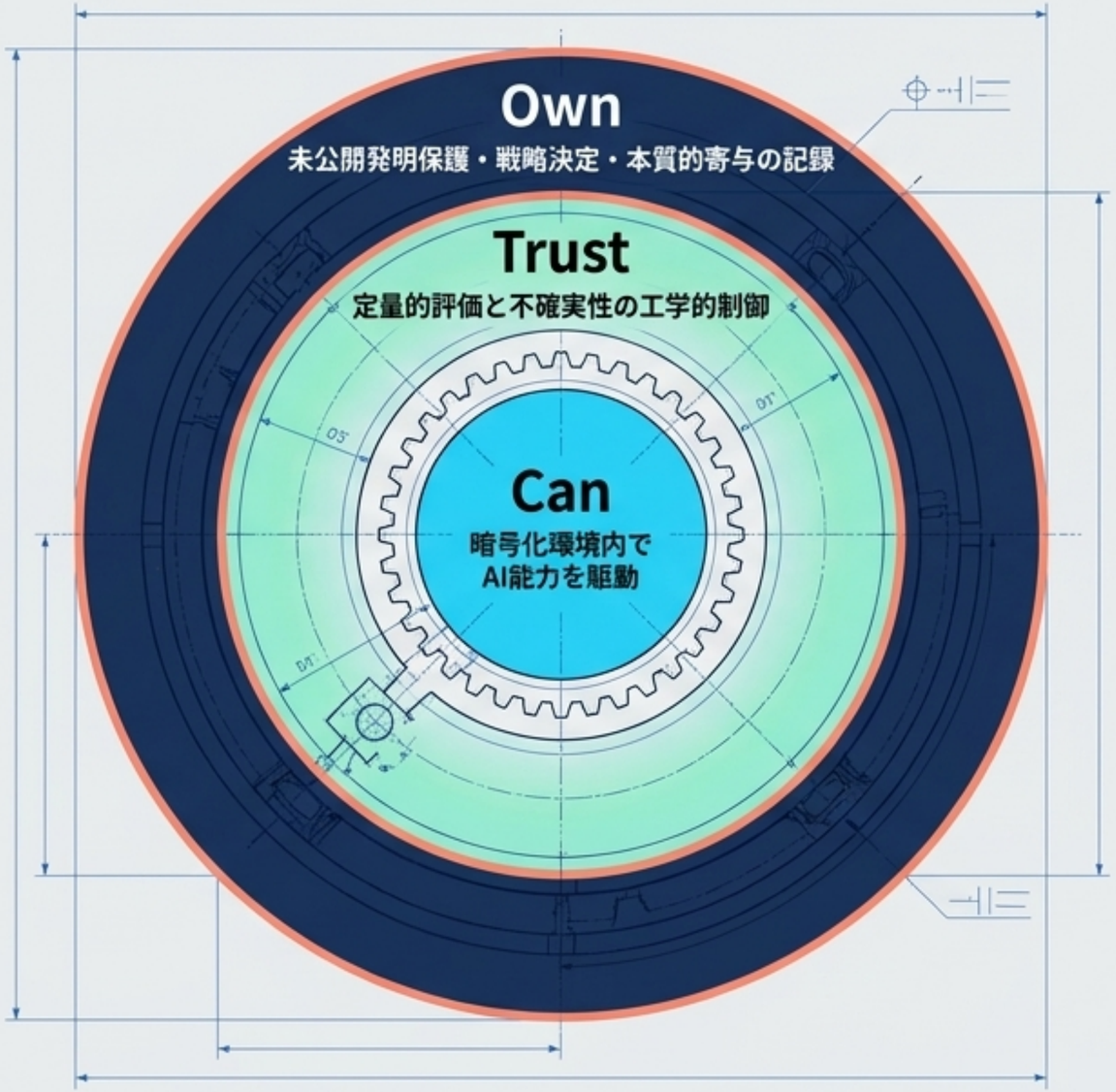
個々の担当者が毎回背景や基準を手入力。
品質のばらつき、資産の非蓄積。

API連携型エージェント（Trust, Own, and Scale）



- 独自の記述スタイル等をRAGとして参照しながら自律的にタスクを遂行。
- 成果: 組織がプロセスを「スキル」としてエンコードし、所有 (Own) し、継続的に改善する体制。

結論：人間とAIの新たな境界線をデザインする



AIは道徳的責任を負えない「予測機械」であり、「Reliance（依存）」の対象に過ぎない。「人間が最後に確認する」という受動的な態度から脱却せよ。

未公開発明の保護や権利化戦略の決定といった法的責任は人間が完全に「Own」し、技術的ガードレール（Trust）の中で言語能力（Can）を駆動させる。

人間が主体的に設計した堅牢な安全基準の内部でのみAIを駆動させること。これこそが次世代知財部門の最も強靭なオペレーティング・モデルである。