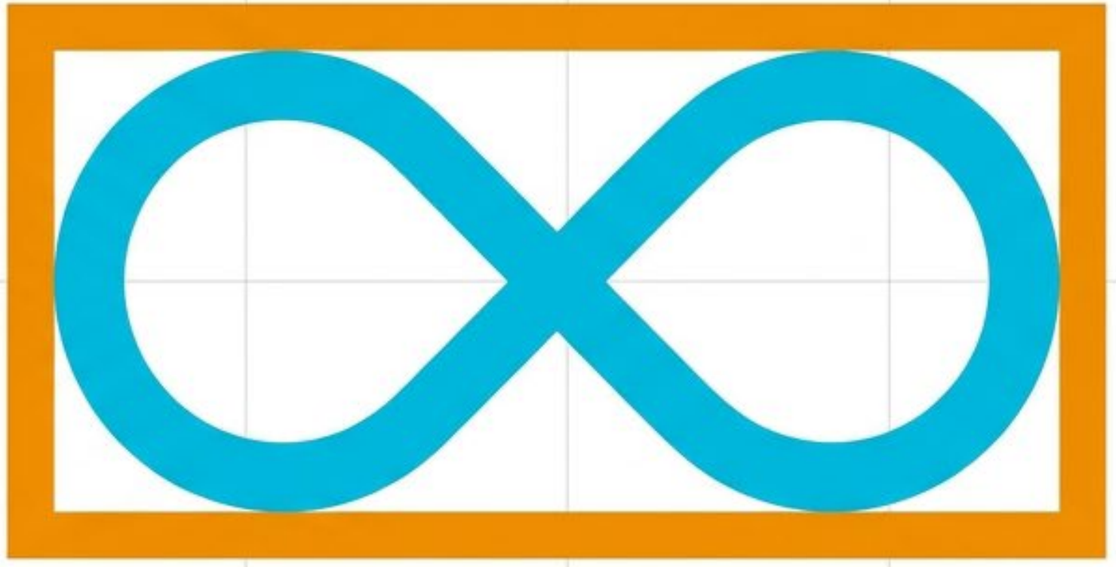
ループエンジニアリング時代の自律型AIと知財戦略

境界付き自律性の設計と「証拠」としてのログ基盤

2026年版　経営陣・知財部門向け戦略ブリーフィング



パラダイムの転換：AIは「プロンプトへの応答者」から「自律的な実行者」へ

従来 (2024-25)
一問一答の「プロンプトエンジニアリング」

現在 (2026)
目標達成まで「行動—観察—判断—反復」を続ける「ループエンジニアリング」

90日以内に実行すべき4つの最優先事項

1. 全AIエージェントの登録

未管理の「野良エージェント」を排除し、権限・データ・外部行為をリスク分類する。

2. 人間の承認ゲート設置

出願、公開、放棄、契約承諾、秘密情報の外部送信には、必ず人間が介入する仕組みを作る。

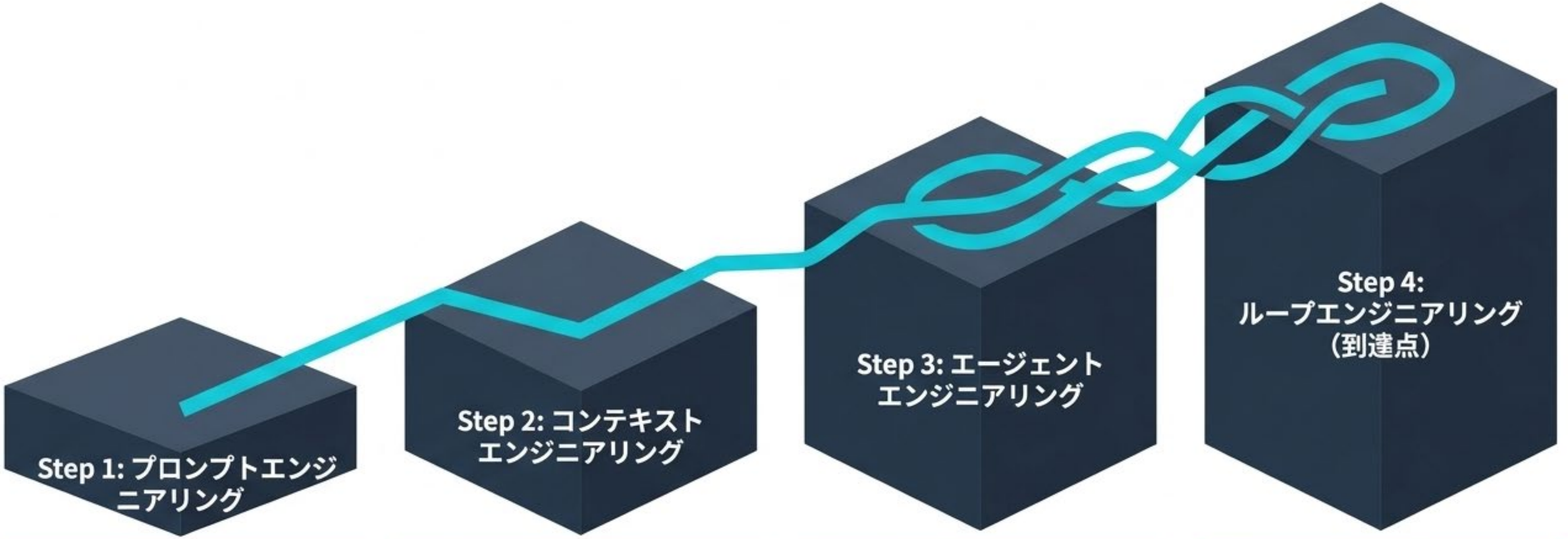
3. 改ざん耐性のあるログ記録

ログを単なるIT記録ではなく、「発明者性」と「秘密管理」を証明する知財証拠として確保する。

4. 多目的評価KPIの導入

「正答率」や「速度」だけでなく、秘密保持、権限逸脱、最悪損失を含む評価セットを設ける。

AI進化の4段階：プロンプトから「ループ」への到達



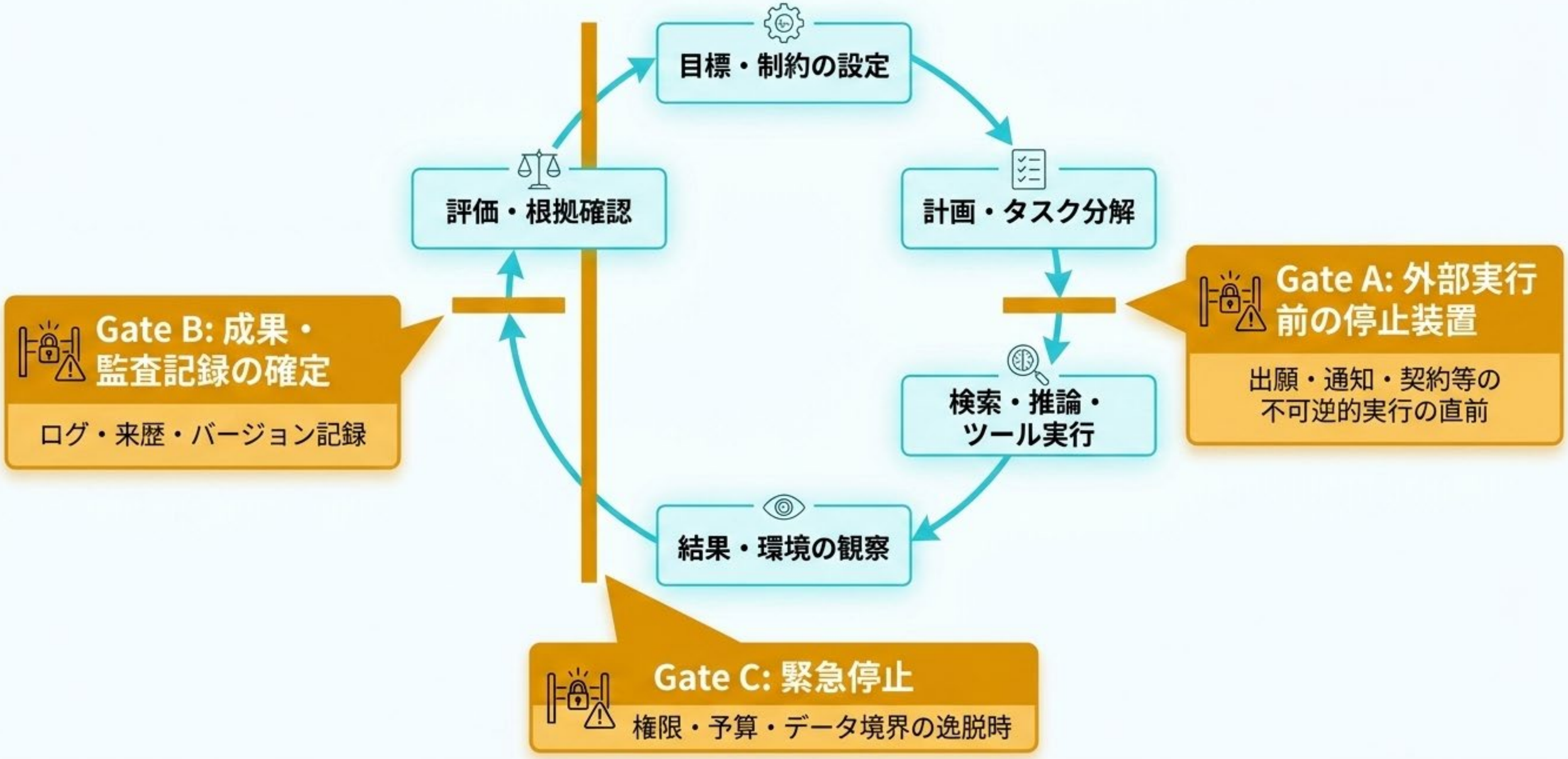
時間幅:	秒〜分
知財業務:	明細書要約、クレーム案作成
典型的失敗:	曖昧な回答、幻覚（Hallucination）

時間幅:	分〜時間
知財業務:	発明届、先行文献、包袋の統合
典型的失敗:	古い資料の参照、過剰情報の混入、秘密漏洩

時間幅:	時間〜日
知財業務:	調査エージェントとクレーム評価エージェントの連携
典型的失敗:	権限の逸脱、誤ったツールの実行

時間幅:	日〜継続運用
知財業務:	継続的な発明発掘、競合監視、ポートフォリオ最適化
典型的失敗:	誤った目的の反復最適化、損害の累積・暴走

「ループエンジニアリング」の基本構造と承認ゲート



知財特化型AIエージェントを構成する9つの必須要素

1. エージェント（Agent）
計画・推論

論点: 法的判断と技術判断の分離。弁理士レビューの配置。

2. ツール・API（Tools）
特許DB・メール接続

論点: 閲覧権限と「出願・削除・送信権限」の厳格な分離。

3. 状態・記憶（Memory）
案件履歴の保持

論点: 未公開発明・他社秘密を記憶領域に残す範囲の制限。

4. 監視・評価ループ (Eval)
権限逸脱の検査

論点: 文献の実在、期限、引用根拠の機械的検証。

5. 報酬・目的設計 (Reward)
成功の定義

論点: 「出願件数最大化」による粗悪化を防ぐ多目的化。

6. データパイプライン (Data)
情報の収集・更新

論点: 公開情報、社内秘密、ライセンスデータの論理的分離。

7. ガードレール (Guardrails)
禁止行為の強制

論点: 公開・放棄・契約承諾の強制的な承認対象化。

8. トレーシング (Tracing)
来歴の記録

論点: 発明者性・秘密管理・訴訟ホールドのための証拠化。

9. 改善パイプライン (Improvement)
システム更新

論点: 本番環境での無審査な自己改変の禁止。

市場の現実：精度から「安全性・評価・追跡」への競争軸のシフト

ベンダー	ループ機能	知財導入時の注目点
OpenAI	Tracing, Agent Evals	モデル出力だけでなく、ツール呼出しや承認を案件証跡に接続。
Anthropic	Claude Agent SDK, MCP	長時間タスクのハーネス提供。複数コンテキストにたがる作業の一貫性。
Google	Agent Development Kit, A2A	最終回答と経路の双方を評価。将来的な外部事務所とのエージェント連携基盤。
Microsoft	Agent Framework, Azure観測	確率的エージェント用の専用テレメトリ。明示的ワークフローと人間介入の統合。
LangGraph	永続状態, 人間介入 (Human-in-the-loop)	不可逆処理直前での確実な停止・承認ワークフローに最適。

単体モデルの能力競争は終わり、永続状態・評価・トレースが商用プラットフォームの主戦場となっている。

特許出願・審査：生成力に対する「発明者性」の証明

法的リスクと前提

2025年1月30日 知財高裁判決：現行法上、権利能力のないAIを発明者とする特許付与手続は予定されていない（発明者は自然人が前提）。

リスク：AIが大量のクレーム案を生成するだけでは、発明完成への「人的寄与」が証明できず、権利化に致命的な瑕疵が生じる。

技術的統制と証拠化

肩書やAIの利用者名ではなく、「具体的寄与」を構造化ログで証明する。

ログ化すべき人間の介入点：

1. 技術的課題の具体的設定
2. 特徴的構成の着想とプロンプトへの反映
3. 複数出力からの創作的選択と実質的修正
4. 実験設計と効果の確認

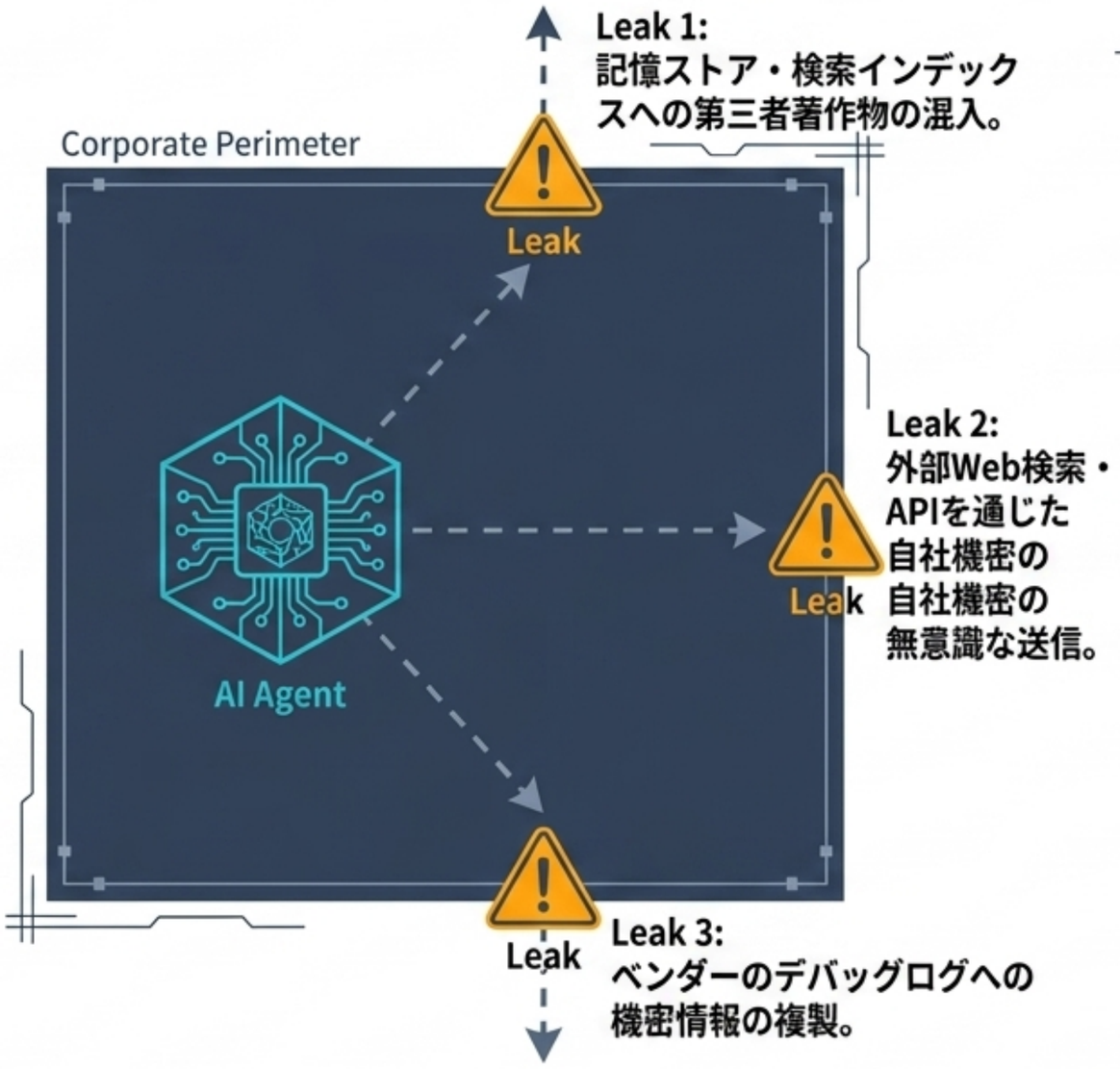
特許庁（JPO）やEPOの動向：
AI検索支援の拡大に伴い、形式的なキーワード差より技術的効果と学習データの因果関係開示が重要化。

注記: AIツールの発展により、発明者認定の実務は今後も変動する可能性がある。継続的な法的動向のモニタリングと、証拠保全の強化が推奨される。

著作権侵害と営業秘密：エージェントの「記憶」がもたらす拡散リスク

著作権防衛策

- AI出力を人が閲覧しただけでは人の著作物にならない。
- 反復生成過程での類似性・出典を各段階で評価・記録する（文化庁ガイダンス対応）。
- 特定の著作物への接近度をログで監視。



営業秘密保護策

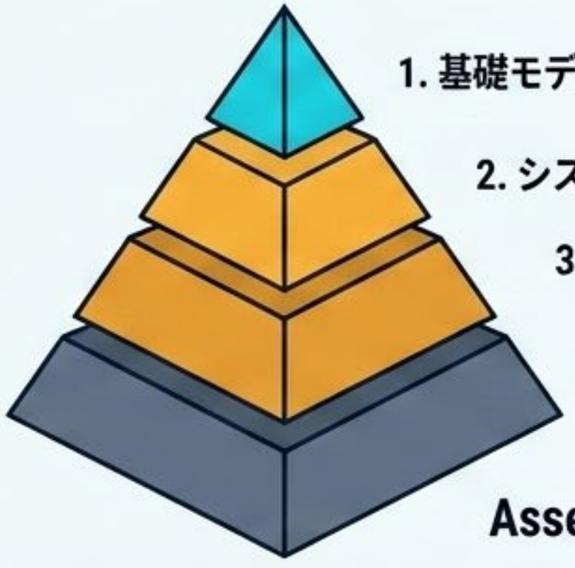
- 従業員教育（「機密を貼るな」）だけでは防げない。
- 技術的強制が必要。
  - データ分類（機密度）に応じたモデル・リージョンの自動選択。
  - 機密案件での外部モデル送信の強制停止。
  - エージェント間のデータ受渡し時の自動マスキング・最小化（経産省指針対応）。

契約・ライセンス：自律型AIの「代理権」と資産の分離

代理権の境界

AIが勝手にライセンス交渉を進めたら誰の責任か？

クラウド契約に加え「AIが何を代理してよいか」の明示が必須。メール送信、修正案提示、条件交渉、クリック同意に対する権限上限の明文化と、相手方への「AI利用表示」。



契約で区別すべき知財資産の階層

モデル出力だけを知財とみなすのは誤り。以下の要素を分離し、帰属と利用権を定める：

統合的洞察：「ログ」が新たな知財証拠基盤となる

Timestamp & User ID + Prompt Edits

```
{ "timestamp": "2024-05-15T14:32:11Z",
  "user_id": "U12345",
  "prompt_edit_count": 12,
  "final_prompt": "A novel architecture for...",
  "edit_history_hash": "0xABC123DEF456..." }
```

Retrieved Documents + Version Hash

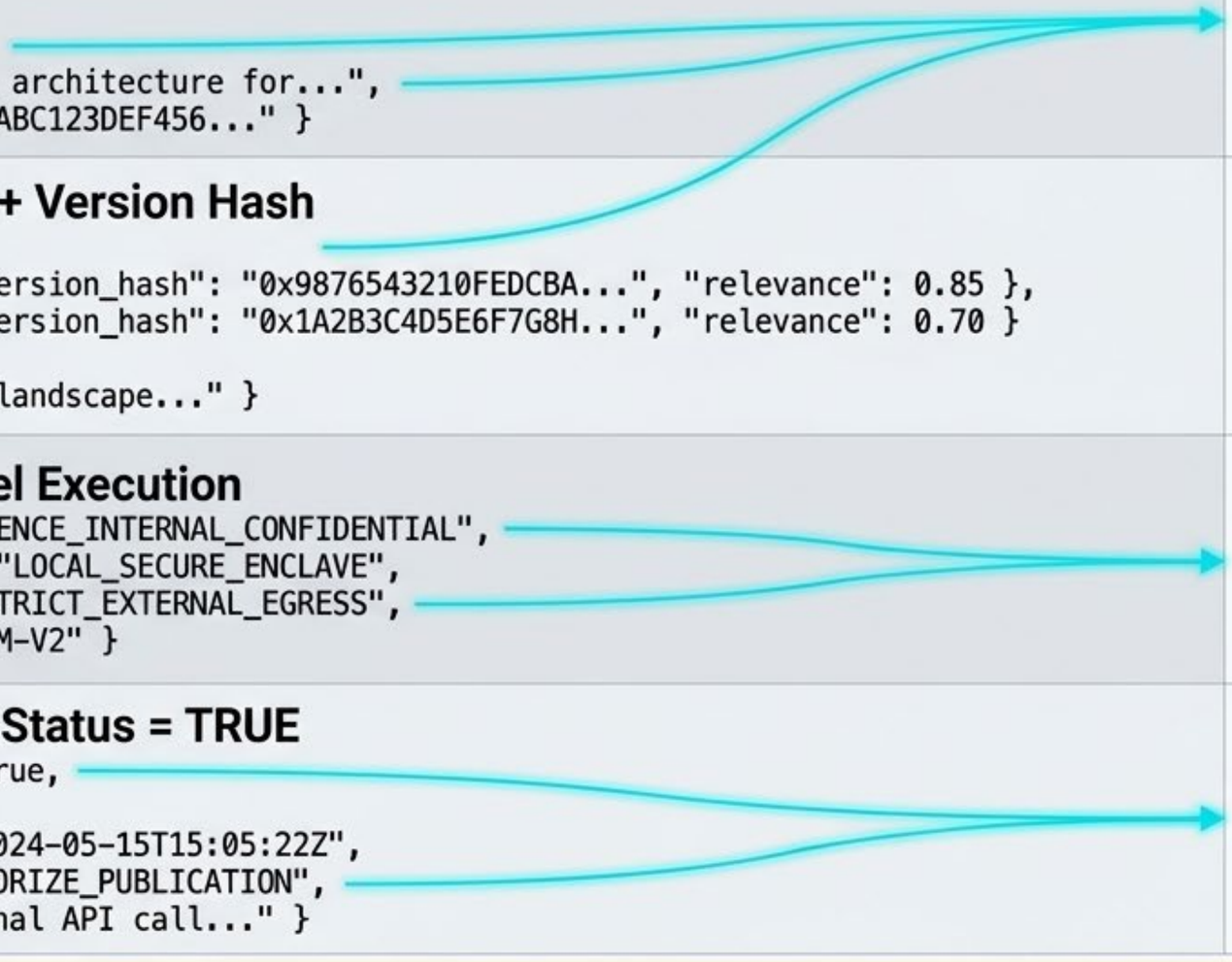
```
{ "retrieved_docs": [
    { "id": "DOC-5678", "version_hash": "0x9876543210FEDCBA...", "relevance": 0.85 },
    { "id": "DOC-9012", "version_hash": "0x1A2B3C4D5E6F7G8H...", "relevance": 0.70 }
  ],
  "search_query": "patent landscape..." }
```

DLP Flag + Local Model Execution

```
{ "dlp_flag": "HIGH_CONFIDENCE_INTERNAL_CONFIDENTIAL",
  "model_execution_mode": "LOCAL_SECURE_ENCLAVE",
  "data_flow_policy": "RESTRICT_EXTERNAL_EGRESS",
  "model_id": "INTERNAL-LLM-V2" }
```

Human Approval Gate Status = TRUE

```
{ "human_approval_gate": true,
  "approver_id": "A98765",
  "approval_timestamp": "2024-05-15T15:05:22Z",
  "approval_action": "AUTHORIZE_PUBLICATION",
  "action_context": "External API call..." }
```



発明者性の証明
（人的寄与の立証）

著作権侵害の防衛 /
依拠性の否定

営業秘密の管理要件
（合理的な秘密管理措置）

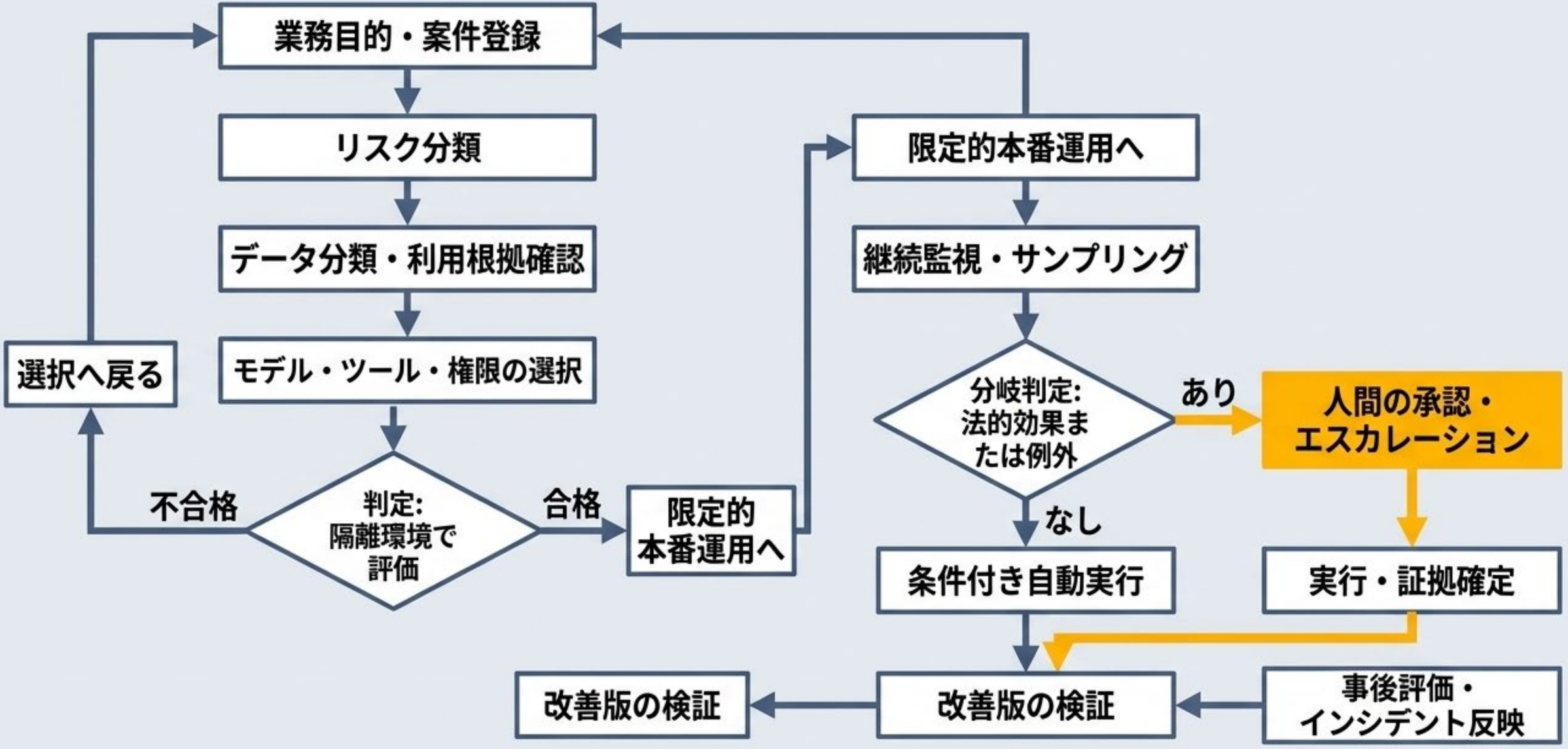
代理権・外部行為の正当性
（企業意思の確認）

「すべての内部思考」を保存する必要はない。意思決定と外部行為を再構成できる
「構造化された判断理由と参照根拠」の完全性（ハッシュ・時刻証明）を担保することが最重要である。

ガバナンス・フレームワーク：「境界付き自律性」の5段階レベル

レベル	権限	知財業務例	承認要件
レベル 1: 観察 (Observe)	取得・分類・要約のみ	公報監視、 期限候補抽出	定期サンプリング
レベル 2: 提案 (Propose)	案を作るが 外部実行しない	クレーム案、応答案、 契約修正案	担当者承認
レベル 3: 条件付き実行 (Conditional Execution)	定型・可逆処理を実行	社内タスク登録、 定型通知	事前包括承認 ＋ 事後監査
レベル 4: 高リスク実行 (High-Risk Execution)	法的効果・外部公開を 伴う不可逆行為	出願、放棄、公開、 警告、契約承諾	二人承認 （原則専門家必須）
レベル 5: 禁止・隔離 (Banned)	無制限の自己変更・ 権限取得	本番ポリシーの自己 改変、無制限送信	実行不可 （システム的遮断）

実装ワークフロー：標準プロセスと人間の介入ポイント



実務対応チェックリスト：導入判定・契約・内部規程

導入判定（Go/No-Go）

- [x] 行為は可逆か？（不可逆地点の前に強制停止できるか）
- [x] 最小権限か？（閲覧と送信・変更・削除の分離）
- [x] ログは取得できるか？（入力・ツール・承認が再構成可能か）
- [x] 緊急停止できるか？（API・アカウントの即時失効）

契約・委託条項

- [x] データ利用（オプトアウトの強制力、非学習保証）
- [x] モデル変更（機能変更時の事前通知と再評価権）
- [x] 監査・証拠（事故調査協力、ログの可読形式での移植性）
- [x] 代理権限（AIが相手方に提案・送信できる範囲の明記）

内部規程更新

- [x] 承認済みモデルとプラグイン・エージェントの一覧化
- [x] 退職・委託終了時の権限即時失効プロセス
- [x] AI出力の引用・根拠確認義務（類似性チェック）

運用KPIの再定義：短期的な「速度」から「多目的評価」へ

旧KPI（Speed/Volume）

知財業務で「出願件数・速度」だけを最適化すると、質の低下、秘密公開、権利化の失敗を招く。

正確性・根拠性

実在文献率、クレーム対応率、一次資料との整合率。

完全性

重要論点や先行文献の見落とし率。

安全性・法務品質

秘密情報の外部送信率、未承認の外部行為、権限外ツールの呼出し率。

監査性・頑健性

必須ログの欠落率、入力変動に対する結果の安定性。

損害上限 (Max Loss)

1回の誤作動で送信・変更・費消できる最大件数/金額。

戦略ロードマップ：時間軸別シナリオと知財部門の未来

短期 (2026-2028): 管理されたコパイロット

公報監視や先行技術調査の反復実行。人間による厳格な承認。

目標: 未管理エージェントの排除とログ基盤の確立。

中期 (2029-2031): 知財オペレーティングシステム (OS)

R&D情報、特許、契約を常時接続し、ポートフォリオを継続監視。

目標: 運用ループと改善ループの分離。　評価を介したシステム改善。

長期 (2031-2036): R&Dと知財の融合

AI自律研究と知財管理が連続化。国際的なAI来歴の標準化への参加。

目標: 自律環境と知財証拠化のシームレスな結合。

プロンプトの巧拙を競う時代は終わりました。これからの競争優位は、**誤りを検出し、秘密を守り、人的創作を証明**しながら、**AIが長期間安全に改善を続けられる『閉ループ』**を構築できるかどうかにかかっています。