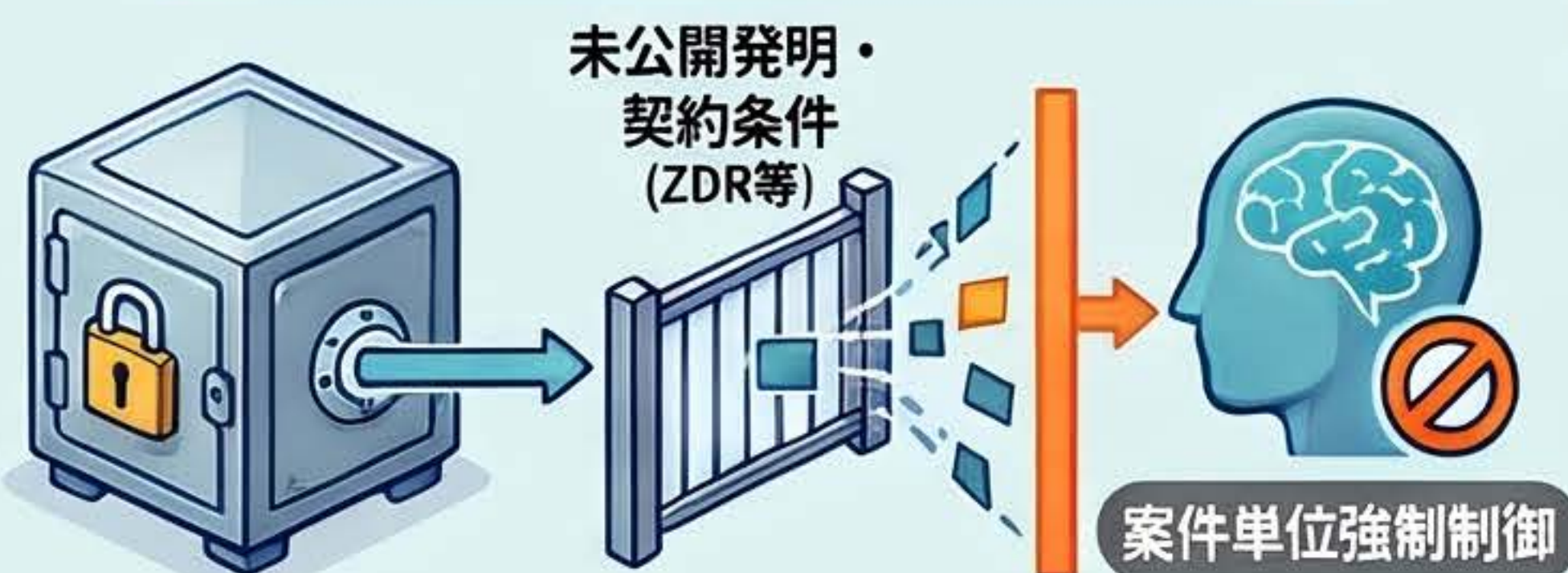


知財AIエージェント運用の要：構造で縛る「ハーネス設計」の鉄則

知財業務へAIエージェントを安全かつ効果的に導入するため、プロンプトに頼らず「構造（ハーネス）」で強制すべき制御原則と承認権限モデル。

ハーネスで決定論的に強制すべき3大ガードレール

機密境界の物理的遮断



未公開發明を渡すモデルや契約条件（ZDR等）を案件単位で強制制御する。

一次DBによる出典の機械照合



特許番号や引用文を公式DBと決定論的に照合するまで出力をブロックする。

不可逆操作における人間の承認ゲート



出願・応答・外部送信など取り返しのつかない操作は人間の承認を必須化する。

操作リスクに応じた4段階の承認制御モデル

L0
L1

L0～L1：閲覧・ドラフト生成（自動 / 検証表示）
公開情報の検索や下書き作成は自動化し、「未検証」フラグを付与する。



未検証フラグ必須

標準設定：
自動実行（「未検証」選かし・フラグを必須化）

L2

L2：外部影響を伴う操作（都度承認 / ask）
メール送信や外部検索は、送信先等の確認と人間の都度承認を求める。



標準設定：
都度承認 (ask・ホワイトリスト管理)

L3

L3：不可逆な手続・更新（原則拒否 / deny）
出願・応答・期限更新等はAI自律実行を禁止し、弁理士の二重承認を課す。



標準設定：
原則拒否 (deny・弁理士等による二重承認)
弁理士 弁理士