

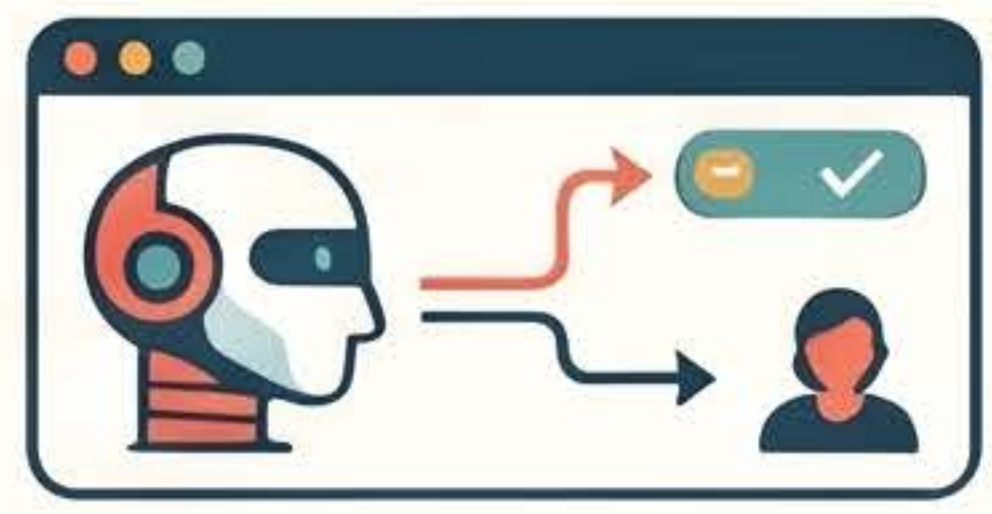
AIへの「判断移譲」：新時代におけるパラダイムシフトとガバナンス要件

判断移譲の3段階パラダイムシフト



「推薦」から
「メタ判断」への拡大

候補提示から、人間へ確認
すべきかどうかの判定（メ
タ判断）までAI側へ移行。



自動承認による人間介
入の削減（200分の1）

AIエージェントの自動承認
レビューにより、人間による
手動停止頻度が大激減。



形式的「Human-in-
the-loop」の限界

確認件数の増大は承認の形骸
化を招き、人間を置くだけでは
安全を担保できない。

次世代AIガバナンスの必須要件



被害度・不可逆性に応じた
「境界付き自律化」

低リスクはAIが自動処理し、
高リスク・不可逆な案件のみ
人間や外部制御層へ戻す。



AI自身が解除できない
「独立した安全境界」

サンドボックスやAPI上限な
ど、モデルの言語的認識に左
右されない制約を配置する。

	主要リスク種別	主要要因	対策の方向性
	Automation bias (人間の確認形骸化)	確認疲れ・ 処理速度の差	不確実性に基づく自動棄却 (Abstention) の組み込み
	Prompt Injection × 過剰権限	判断権限と 実行権限の結合	最小権限原則・サンドボックス等の 独立制御
	責任主体の 不透明化	責任帰属の 曖昧化	開発・提供・利用の多層的責任分配 とログ保存