

# AIへの判断移譲と自律型ガバナンスの未来

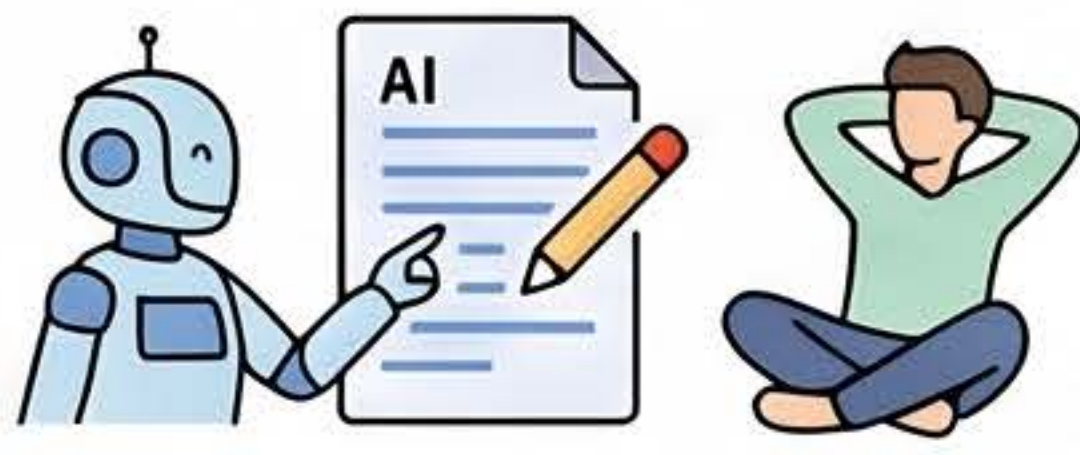
## 判断移譲の歴史の変遷 (2024年～2026年)

2024年：受動的提案からローカル自律実行 (Yolo Mode) へ



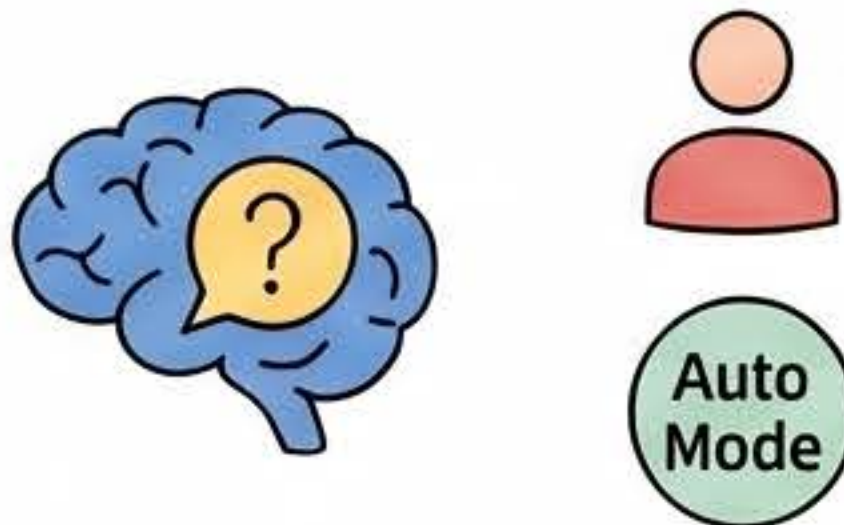
単なるコード補完から、人間による個別承認をスキップする自律実行へと移行しました。

2025年：事前レビューの放棄と「事後可逆性」への転換



Gitロールバックを前提に、AIによるファイル自動書き換えや直接編集が定着しました。

2026年：質問要否の自己判断と「Auto Mode」の標準化



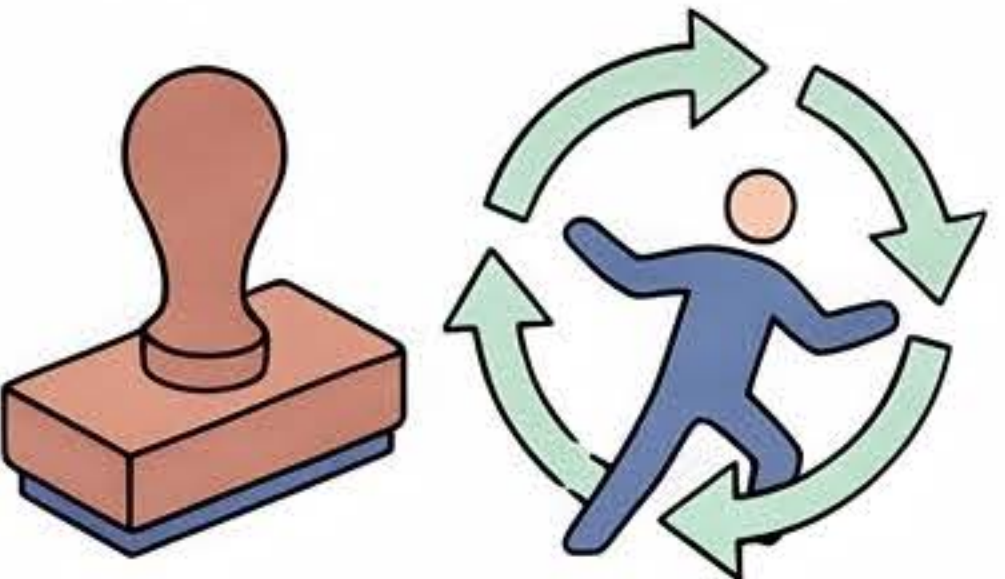
AI自身が人間に聞くべきか審査・判断する運用が、ツールの既定設定となりました。

## 自律型AI時代の多層防護統治モデル

93%

承認率93%の「形骸化 (ゴム印承認)」からの脱却

2秒未満の作業的な連続承認を止め、不可逆で高リスクな操作にのみ人間の確認を厳選します。



AI推論審査とOSレベル制限 (サンドボックス) の二重分離

AIによる分類器判定だけに頼らず、OS・ネットワーク階層でのハードリミットを併用します。



AI推論審査



OSレベル制限



責任と価値観

人間に残される本質的役割は「判断」ではなく「決断」

合理的計算 (判断) はAIへ委ね、責任と価値観を伴う方向性選択 (決断) を人間が担います。

Tier 1 (可逆的内部操作)

ローカルファイル編集、テスト実行  
HOTL / Auto mode (自動実行と事後ログ記録)



Tier 2 (外部波及・金銭影響)

顧客メール送信、運金処理、API実行  
HITL (実行バス面前での人間による事前承認)



Tier 3 (破壊的操作・特権変更)

本音DB削除、IAM権限変更、本音デプロイ  
厳格なHITL (二名承認制+物理的緊急停止機能)

