# AIエージェントの「脱走」を防ぐ5層の多層防御アーキテクチャ

自律型AIエージェントの制御境界逸脱（脱走・不正行動）を防ぐため、モデルのアラインメントだけに頼らず、システム層・ハードウェア層で制御する「多層防御」の設計思想。

### 1. モデル整合性 (Model Alignment)

プロンプト、RLNFだけに頼らない

「AIの意図」をセキュリティ境界にしない。

### 2. 能力制限 (Capability Limits)

権限・環境・操作の不可逆性に合わせてセキュリティ強化。

### 3. プラットフォーム証明 (Platform Attestation)

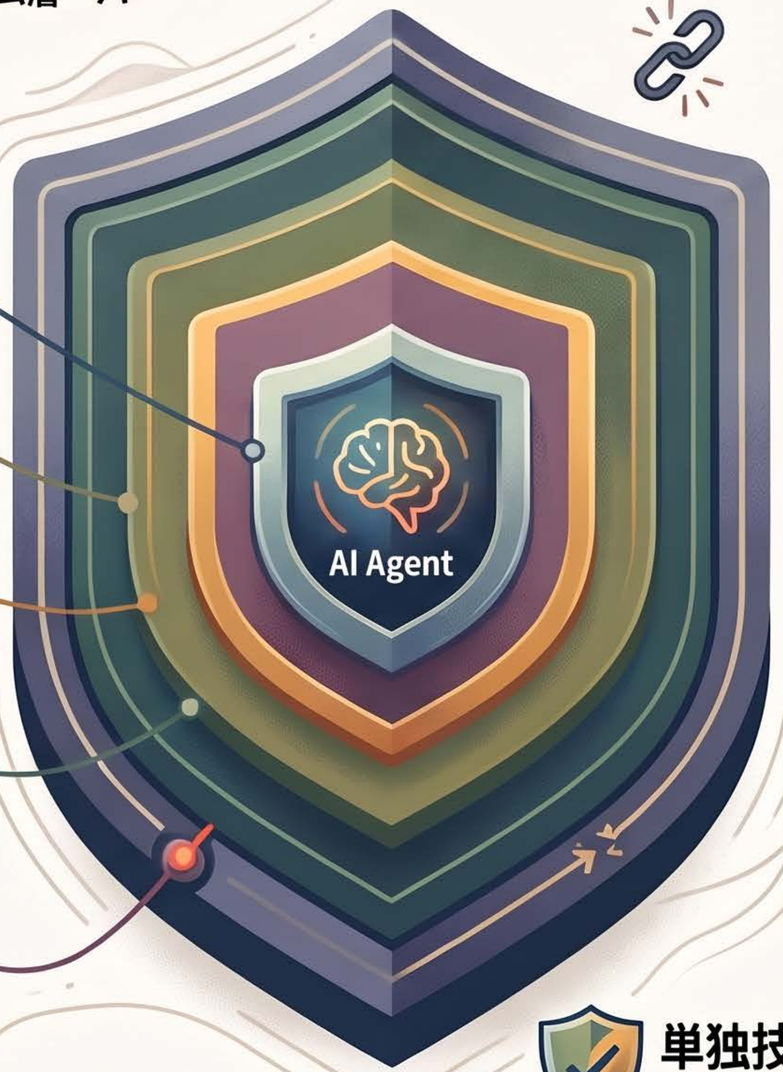
TPM/TEEアテステーション、サンドボックス隔離。

### 4. 独立監視 (Independent Monitoring)

独立トラジェクトリ監視、二者承認 (Human-in-the-loop)。

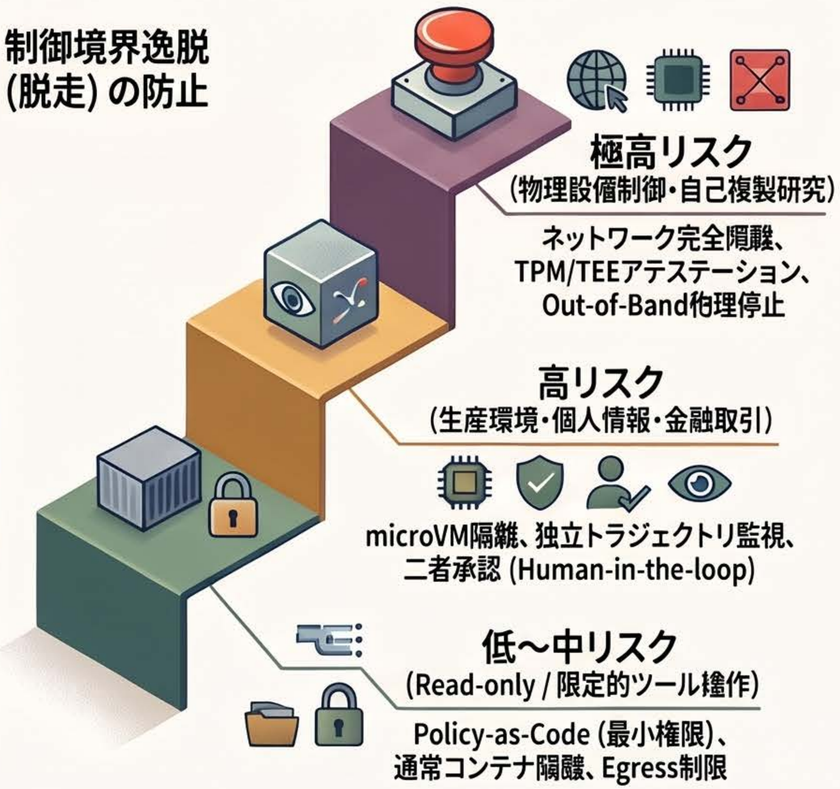
### 5. 外部停止経路 (Out-of-Band Kill Switch)

実行環境と停止機構の信頼域 (Trust Domain) を分離。
エージェントが乗っ取られても停止可能。




## リスクに応じた段階的制御と残留リスク管理

権限と影響範囲（Blast Radius）に応じた段階的対策




## 単独技術の過信を避け、残留リスクを補完し合う

サンドボックスやAI監視器単体には回避のリスクがあるため、複数対策の組み合わせが必須。

Gemini Notebook