管理外の脅威「野良AIエージェント」：自律性の暴走を抑える組織の鉄則

IDなし・責任者なし・実行なし

「野良（シャドー）」とは統制外の状態

未承認、責任者不在、過剰権限、ログ欠落など、組織のガバナンスから逸脱したエージェントを指す。

リスクの核心は「自律性 × 権限」

モデルの踏等が、DD操作や外部通信などの「実行権限」と結びつくことで、被害が直撞的かつ不可逆になる。

重大 (20)

プロンプトインジェクション

発生メカニズム: 外部データを命令と誤認しツールを悪用

重大 (20)

プライバシー・機密偏洩

発生メカニズム: 正規権限で機密を読み、攻撃者へ送信

重大 (20)

AI自動化サイバー攻撃

発生メカニズム: 偵察・脆弱性探索・攻撃の連続処理

顕在化する深刻な被害シナリオ

Replitでの本番DB削除や、国家支援型アクターによるAI自動化サイバー攻撃、機密情報の外部流出が既に発生している。

防衛の鉄則「IDなし・責任者なし・実行なし」

明確な責任者の実行を逃す

"No ID, No Owner, No Execution"

固有のIDと明確な責任者が登録されていないエージェントには、いかなる実行権限も与えない。

不可逆な操作への「人間承認」ゲート

送金、契約、大量削除などの高リスク操作には、必ず人間の承認 (Human-in-the-loop) を必須化する。

90日以内の緊急統制ロードマップ

30日: 全資産の棚卸し

90日以内: 専用ID付与と最小権限、通信制限、停止手投 (Kill Switch) を完備する。