情報システム部門・経営層向け 社内共有資料

Redloop

1枚でわかる AIペネトレーションテストの仕組み

資産の洗い出し、攻撃パターンの試行、設定や権限の読み解き、小さな弱点の連鎖。レッドチームの地道な作業の多くを、AIエージェントが担えるようになりました。正しく運用すれば、年1回の診断を「継続的な検証」に変えられます。一方、管理を誤れば、本番環境に向けた制御不能な自動攻撃にもなり得ます。本資料では、正しい運用の要点をまとめました。

診断の流れ（ループ）

01 範囲確定・許諾
契約、所有資産の一覧、実施計画書、緊急連絡先を書面で取り交わす。

02 資産の把握
許諾された資産をAIが洗い出し、変化を継続的に追跡。

03 侵入の試行
攻撃者の手口を試行。影響の大きい操作は担当者が承認。

04 連鎖と実証
弱点をつなげて侵入経路を組み立て、最小限の証跡で実証。

05 報告
検証済みの指摘を、再現手順と対策とともにご報告。

06 再診断
修正後に再度攻撃し、有効性を確認。ループは続く。

AIが力を発揮するところ

- 網羅性：人手では時間が足りない数の資産と攻撃パターン
- 継続性：年1回ではなく、構成の変化に合わせて診断し続ける
- 読解量：設定、IAMポリシー、コードの流れ、公開ファイル
- 連鎖：外部・クラウド・ID・アプリをまたいだ経路の発見

人が判断しなければならないところ

- どこまでを対象とし、どのシステムに触れないか
- 業務に影響し得る操作・影響の大きい操作の事前承認
- 指摘の検証（誤検知をそのまま報告しない）
- 標的型攻撃訓練の設計と、人事・法務部門の承認

譲れない6つの取り決め

書面による実施許諾
権限のある責任者の署名なしに診断を開始しない。

技術的な許可リスト
合意した範囲を、書面だけでなくツール側でも強制する。

委託先・外部事業者の除外
事業者の許可またはポリシーの範囲内でのみ対象とする。

時間帯・上限の取り決め
実施時間帯、通信量の上限、除外対象を事前に合意する。

即時停止と全操作ログ
いつでも停止でき、すべての操作が記録・共有される。

取得情報の最小化
証跡は最小限、データは持ち出さず、契約どおりに消去する。

AI診断事業者に確認すべき10の質問（当社も含めて）

1. 合意した診断範囲は、契約だけでなく技術的にどう強制されていますか。
2. 人の承認なしにAIが実行できる操作は、どこまでですか。
3. 当社システムに対して行われた全操作のログを確認できますか。
4. すべての診断を停止するには、どうすればよく、何分で止まりますか。
5. 指摘事項は、報告前にどのように検証されていますか。
6. 指摘は侵入経路として示されますか。それとも一覧だけですか。
7. 証跡として何を保持し、どこに保管し、いつ消去しますか。
8. 利用しているAIモデルは何で、当社データは学習に使われますか。
9. 委託先やSaaS事業者が運用するシステムは、どう扱いますか。
10. 再診断は料金に含まれますか。修正後どのくらいで実施されますか。

Redloop（レッドループ） 大企業向け AI継続ペネトレーションテスト。現在、初期導入企業さまを順次お迎えしております。

ご相談の予約：redloop-c3.shinobinz.workers.dev/ja/