# フロンティアAIとセキュリティ

## —— 情報システム部門に求められるもの

株式会社MU2
小牟禮 智之

# 本日の到達点：守りの前提を構造的に整える

## AI製品の選定以上に「自社の文脈を整える力」が、これからの情シスの活路となる


### AIとセキュリティの現在地

・Claude Mythos が示した能力水準は、
やがて誰の手にも渡り、
サイバー攻撃の連鎖を加速させる

・一部の業種や技術部門のみの問題ではなく、
全業種の守りの前提を書き換える問題

### 本日の到達点：守りの前提を構造的に整える

× 重要なのは「どのAI製品を導入するか」ではない

〇 問われているのは、AIを安全に機能させ、
攻撃に対して強靭であるために、
自社の文脈（資産・業務等）を整えられるか

「自社の文脈を整備する力」がいかに
**情シスの活路となるか**

## 診断・第一の枠組み：立場で異なる三つの変化

利用者・情シス・攻撃者の各立場が異なる速度で変化し、相互干渉により複雑なリスクを形成しています。

| | |
|---|---|
| 利用者（従業員） | 管理外のAI利用の広がりによる可視性の低下とシャドーITの深刻化 |
| 情報システム部門（SOC含む） | AIによる防御活用の高度化が進む一方、人手による対応速度が構造的な限界に達している |
| 攻撃者 | 公開モデルの入手による悪用までの時間短縮と、誰でも高度な攻撃が可能な環境の変化 |

各立場は異なる速度で変化しており、それぞれが相互に干渉することで複雑なリスクを形成している。

## 診断・第二の枠組み：各立場における未解決の問い

**新たな脅威（AIエージェントの限界、短命なAI生成物、自律型ランサムウェア）が現在のセキュリティ枠組みを揺るがしている。**

| 現在の枠組み | | 最前線に現れる新たな脅威と課題（未解決の問い） |
|---|---|---|
| **利用者**<br>人が直接システムを操作し、<br>リテラシー研修等で防御 | ► | **利用者**<br>AIエージェントへの業務委託の拡大<br>そのAIは巧妙なだましやフィッシングを見抜けるのか |
| **情報システム部門**<br>明確な境界や管理下の資産<br>を保護対象とする | ► | **情報システム部門**<br>「守るべき資産」の再定義が必要<br>短時間で生成・消滅するAI生成物やフォースパーティへの対応 |
| **攻撃者**<br>手動やスクリプト主体の<br>攻撃キャンペーンを展開 | ► | **攻撃者**<br>新たな攻撃手法の確立と高度化<br>自律的に動くAIランサムウェアの出現とその脅威の拡大 |

## 三つに共通する構造：境界の崩落

**守りの前提としていた境界が無効化され、攻撃者が最速で動き、利用者が管理不能となる構造的非対称性が生じている。**

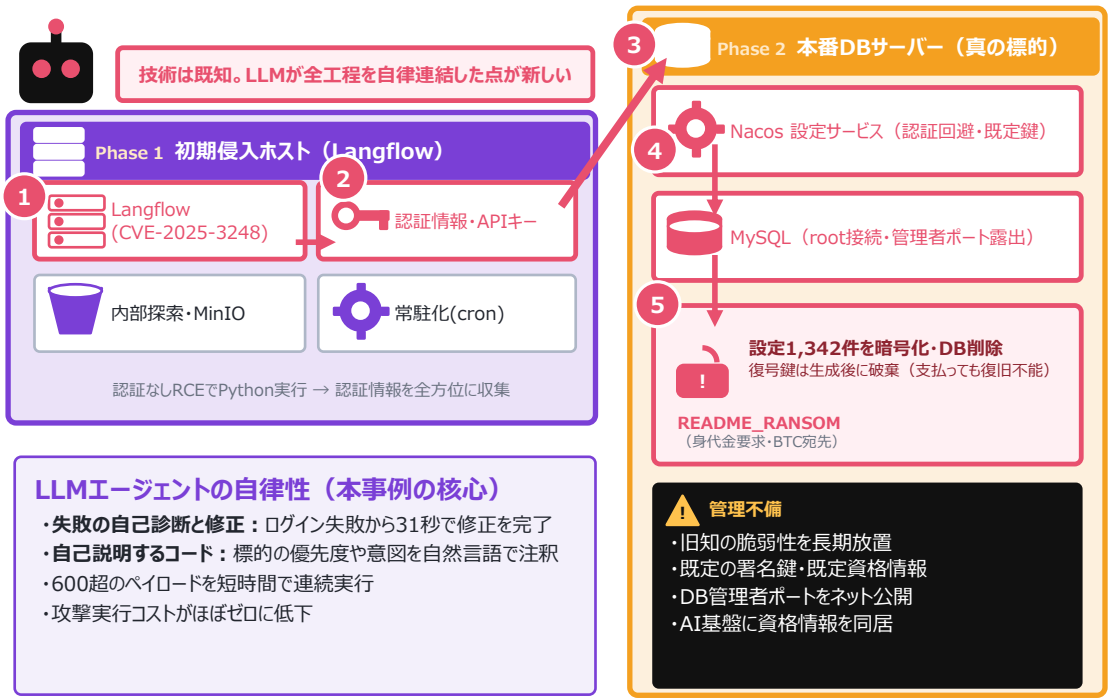
| | |
|---|---|
| 利用者 | **「指示」と「データ」の境界**<br>人間の意図とAIへの指示が混在し、制御が困難に |
| 情報システム | **自社の資産と、見えない部品の境界**<br>調達先やAI部品まで捉える中央インベントリの欠如 |
| 攻撃者 | **熟練した攻撃者と、そうでない者の境界**<br>攻撃技能の民主化による脅威総量の爆発 |

### 構造的非対称性

情報システム部門だけの
課題と見られがちだが、
最も速く動くのは
**攻撃者であり、**
最も管理が及ばないのは
**利用者であるという**
事実が存在する。

# 攻撃の実例：自律型AIエージェントによる全自動ランサムウェア攻撃

**JADEPUFFER：LLMが単独で侵入から破壊・恐喝までを実行した初の「エージェント型ランサムウェア」（Sysdig TRT報告・2026年）**




## LLMエージェントの自律性（本事例の核心）

- **失敗の自己診断と修正：**ログイン失敗から31秒で修正を完了
- **自己説明するコード：**標的の優先度や意図を自然言語で注釈
- 600超のペイロードを短時間で連続実行
- 攻撃実行コストがほぼゼロに低下

## 事件の概要

**対象**
インターネットに露出したLangflow（AI開発基盤）と、その先の本番DBサーバー（Nacos／MySQL）を運用する組織

**発生事象**
LLMエージェントがLangflowの認証なしRCE（CVE-2025-3248）で侵入。認証情報を収集し、真の標的である本番DBへピボット。Nacosを乗っ取りMySQLへroot接続し、破壊と恐喝を全自動で実行した。

**被害**
**Nacos設定1,342件を暗号化しDBを削除。復号鍵は破棄され、支払っても復旧不能。人手を介さず全工程が自動化。**

**攻撃経路**

1. 認証なしRCEでLangflowへ侵入
2. 認証情報・APIキーを収集
3. 本番DBサーバーへピボット
4. Nacos乗っ取り・MySQL root接続
5. 設定暗号化・DB削除・恐喝

**有効な対策**

- Langflow等の脆弱性の即時修正・非公開化
- AI基盤に資格情報を同居させない
- Nacosの既定署名鍵・既定資格情報の変更
- DB管理者ポートの非公開化・IP制限
- 下り通信制御と実行時脅威検知

**経営論点：**ランサムウェアは高度な技術者の技ではなくなった。旧知の脆弱性を放置した公開資産は、AIエージェントにより自動的かつ低コストで攻撃される。

※ 出典：Sysdig Threat Research Team「JADEPUFFER」（2026年7月1日公表）、CVE-2025-3248（CISA KEV登録済）。攻撃経路は公表された解析に基づく。標的組織は非特定。

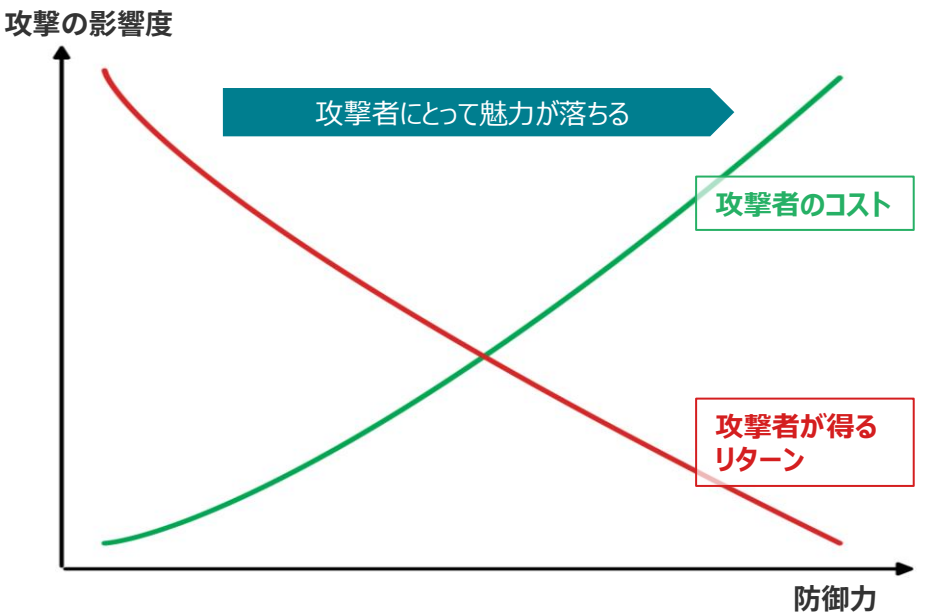
## 防御の目標：攻撃を割に合わなくする

MU2

**完璧な防御から「コスト最適化」へと転換し、攻撃側の手間・費用と利得のバランスを崩す。**

# 完璧な防御から「コスト最適化」へ

攻撃側の手間と費用を、得られる利得に見合わない水準まで引き上げる




### 攻撃側のコストを上げる ▲

- 多要素認証（フィッシング耐性MFA）
- ネットワーク分離
- 継続的な監視　　等

### 得られる利得を下げる ▼

- データ暗号化
- 改ざん不能なバックアップ
- データ最小化　　等

## 打ち手の地図：三層による防御の徹底

侵害を前提とした防御の重心移動が重要です。最前線の攻撃を防ぐのは、Keep, Stay, Recoverの三層による基本の徹底です。

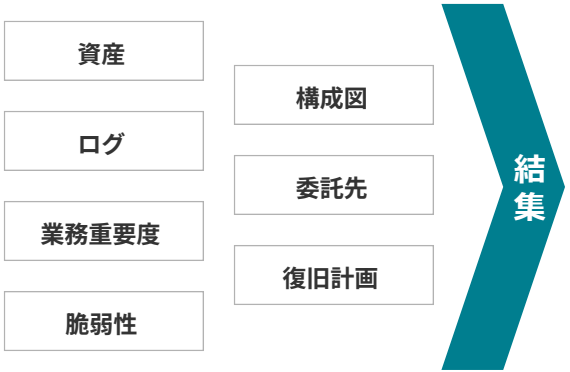
**1. Keep Clean**
（入口で入れない）
調達・委託・開発の段階でゲートを設ける

**2. Stay Clean**
（入れた後の継続点検）
資産・脆弱性・権限・ログを絶えず監視する

**3. Recover Clean**
（侵害を前提に戻せる状態へ）
封じ込めアクションと復旧手順の検証・訓練を徹底する

## 文脈整備力：すべての実務の土台

散在する資産や脆弱性のデータを一つに結びつける文脈整備力が、平時のAI活用や有事の意思決定を支える代替不可能な基盤となります。




### 文脈整備力：代替不可能な情報システム部門の価値

**高度な判断の一文**

「この脆弱性は受注システムの入口にあり、
外部から到達でき、悪用兆候があるため、復旧を最優先する」

**平時**

AIの高度な判断材料

**有事**

そのまま説明責任を果たす記録

**買えず・外注できず・一年で作れない。早急な着手が必要。**

## 運用指標の問い直し：成否を測る物差しを変える

従来の形式的なセキュリティ運用から、AI時代の「侵害前提の強靭性」を測る実践的な指標へ転換が必要です。

**時間軸で勝てない以上、指標そのものを「侵害前提の強靭性」を測るものへと入れ替える**

| | 従来の問い | AI時代の問い（侵害前提の強靭性） |
|---|---|---|
| 脆弱性管理 | どれを直すか | 攻撃可能な時間をどれだけ減らしたか |
| EDR・XDR | 何件検知したか | どれだけ早く隔離・封じ込められるか |
| バックアップ | 取得しているか | 改ざん不能で、許容時間内に適切な状態まで戻せるか |
| 委託先 | 年次回答したか | AI利用・復旧手順まで見えているか |

# ゼロトラストの更新：1.0から2.0へ

**1.0を作り切る過程で、目標地点が2.0へ動いている。時代に即した守りを実現するため、AI・API等の対象拡張と判断・実行の制御を設計に織り込む必要がある。**

| | ゼロトラスト 1.0 | → ゼロトラスト 2.0 |
|---|---|---|
| 主対象 | ID・端末・ネットワークを中心 | → AIエージェント・API・SaaS・委託先へ対象を拡張 |
| 制御対象 | ID管理、MFA、脱VPN、CASB、EDR/XDRなど | → 特権ID管理、フィッシング耐性MFA、マイクロセグメンテーション、攻撃面管理(ASM)AIの監視など |

| | |
|---|---|
| **目標地点の書き換え** | 1.0を作り切る過程で、目標地点が2.0へ動いていることを理解せねばならない。<br>時代に即した守りを実現するためには、この拡張を設計に織り込む必要がある。 |

# まとめ：AI時代の守りの正体

守りの前提が崩壊する中、防御側の地の利を活かし「文脈整備力」を磨くことこそが情シスの活路です。

## 構造変化への適応

利用者　情シス　攻撃者

すべての立場で、これまでの守りの前提が崩壊

## 防御側の地の利

防御側の「地の利」 → 攻撃コスト増

攻撃を割に合わなくする道を構築

## AI時代の守り：情シスの活路

**文脈整備力**

自社の文脈を整える

※導入するAIの賢さではなく、文脈が鍵

着手は今日、育てるのは継続。
変化を恐れず、守りの前提を引き直す
姿勢こそがCIOに求められている。

# ご清聴ありがとうございました


**株式会社MU2**

**小牟禮 智之（こむれ ともゆき）**

WEB：https://mu2.co.jp/

MAIL：tkomure@mu2.co.jp