



信頼、セキュリティ、AI

企業に取り組むべきリスクと対策

山下 頼行

Yoriyuki Yamashita

リサーチマネージャー、Network & Security

IDC Japan



AI Agent Layer

アジェンダ

1 信頼の3要素

2 本日考えたい3つの問い

3 提言

信頼の3要素



自動化に対する慎重な姿勢に対応する「信頼」を設計することは、競争優位の重要な源泉となる。

①



自動対応の根拠を説明可能にする設計

そのAIは、行った対応の根拠を説明できるか

②



人とAIの役割分担

自動化と人による判断の役割分担を適切に設計できるか

③

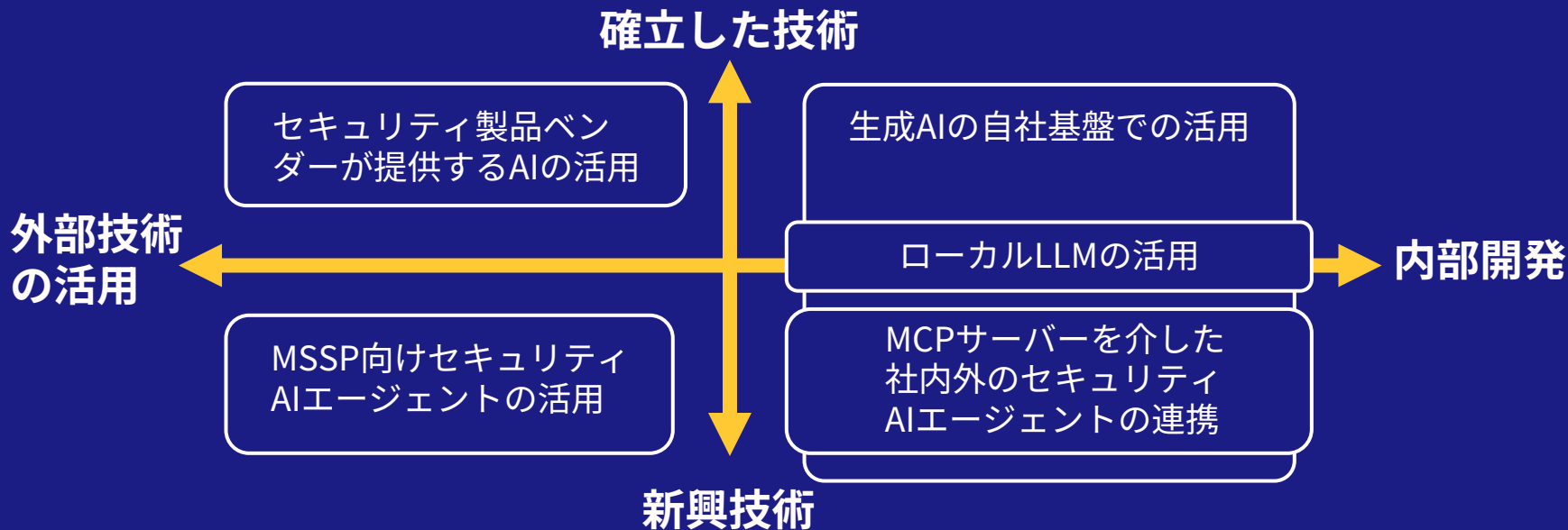


実績の開示

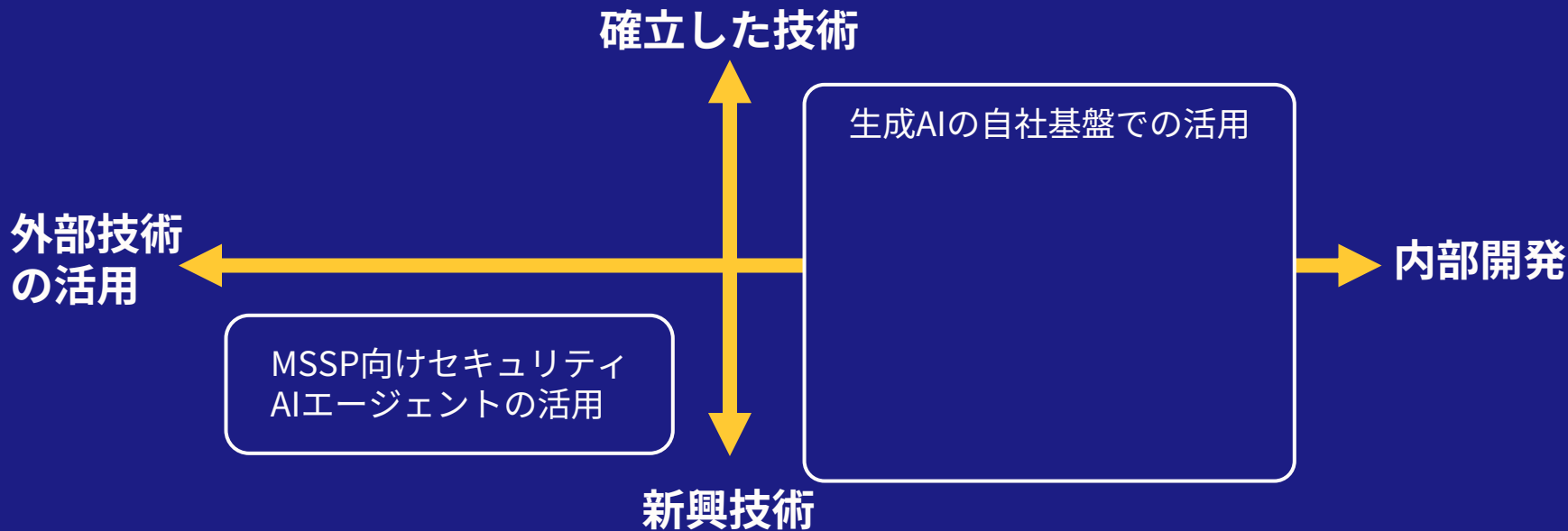
ベンダーは導入事例や検知精度を具体的に示せるか



MSSPによるAIを用いた自動化機能の利用形態

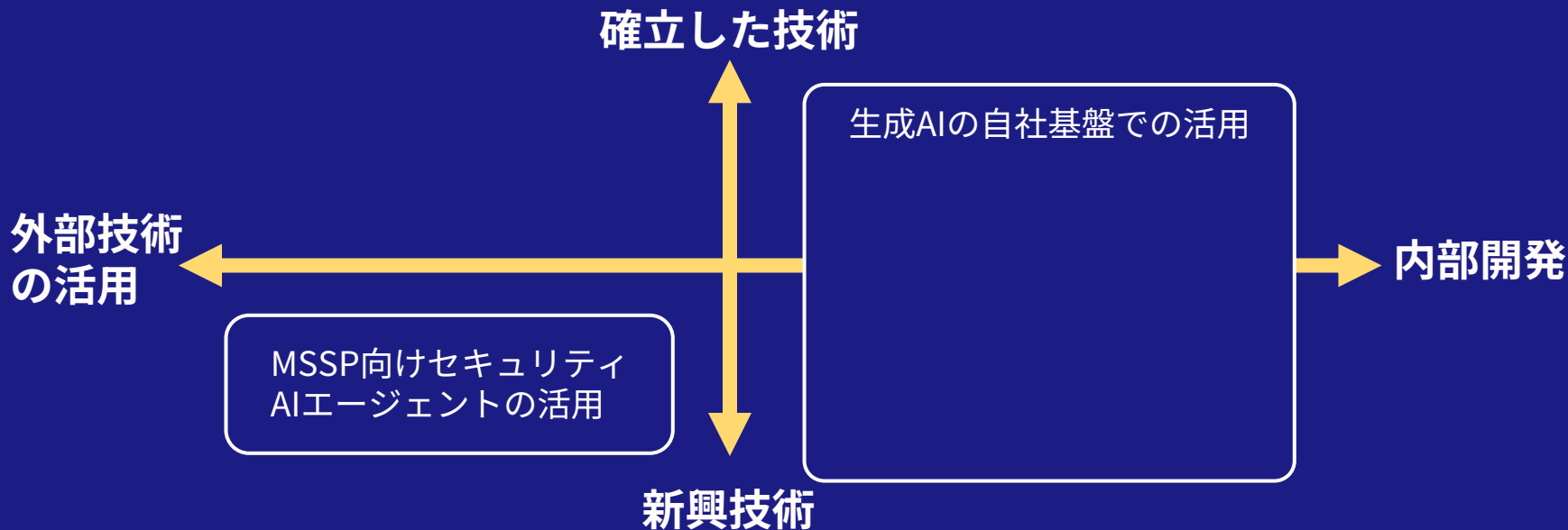


① 自動対応の根拠を説明可能にする設計



ハルシネーション対策としてAI判断の根拠を提示できる設計を採用
推論エンジンで判断根拠を自然言語提示し、経営層報告の品質ばらつきを抑制

②人とAIの役割分担



大部分を自動化するも、人間による判断や確認のプロセスを残す
夜間は自動対応、日中はアナリスト主導で運用



③実績の開示

ベンダーが開示すべき項目の例：

- 数値
 - 誤検知率
 - 平均検知・対応時間
 - 封じ込め成功率
- 顧客事例

数値などの客観資料や、顧客事例を提示することが効果的



本日考えたい3つの問い



本日考えたい3つの問い

1 既存のツール／人間による対応は信頼されているか？
Respond & Recovery

2 生成AI／自動化は信頼されているか？
AI Input & Output

3 AIエージェントの行動は信頼できるか？
Agent Trust

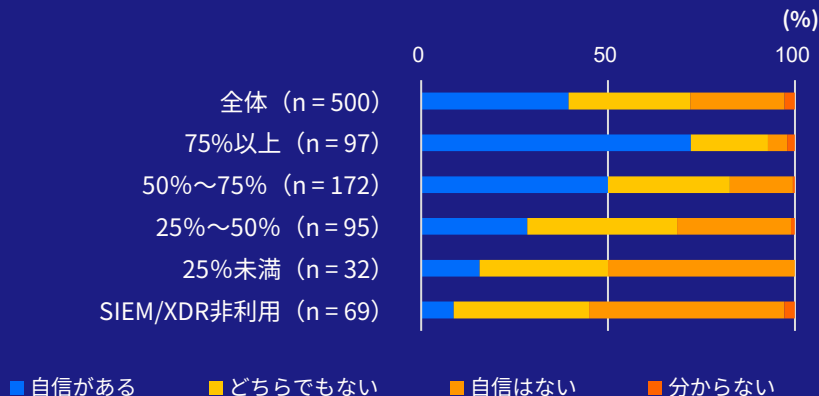


既存のツール／人間による対応は信頼されているか？

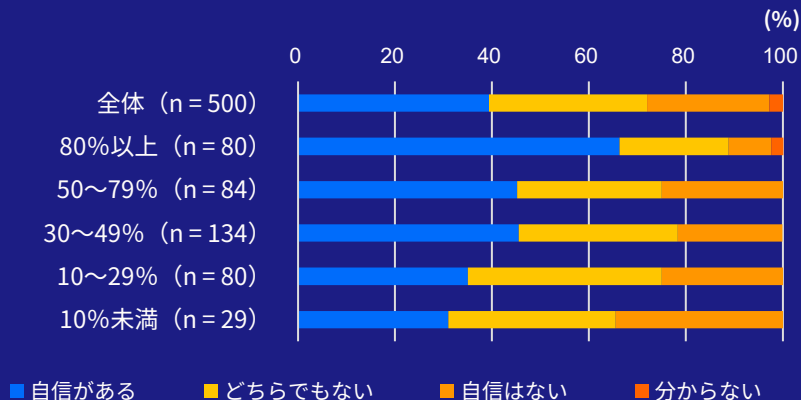


既存のツールは信頼されている

SIEM/XDRへのデータの集約度と 24時間以内に対応する自信



セキュリティプラットフォームへの 機能統合度と24時間以内に対応する自信

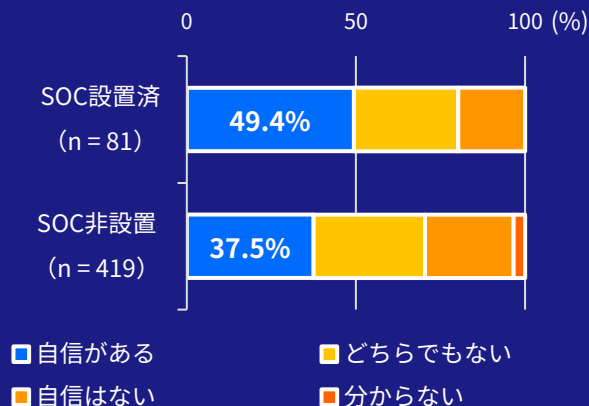


SIEM/XDRへのデータ集約度が高いほど、またセキュリティプラットフォームへの機能統合度が高まるほど、24時間以内に脅威に対応する自信が高まる

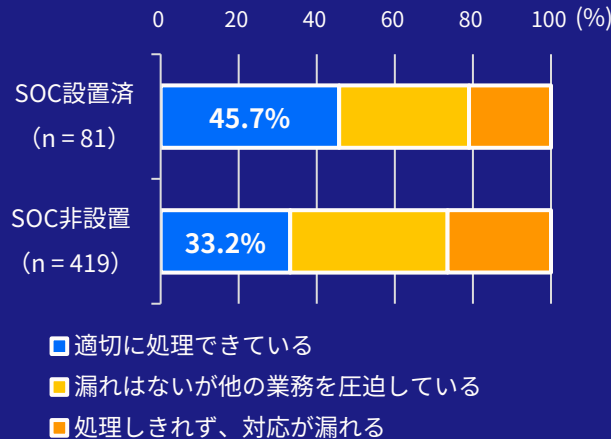


人間による対応は信頼され、効果も挙げているが、利用できる企業は少ない

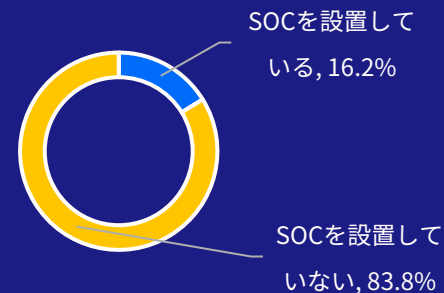
SOC設置有無と24時間以内の対応の自信



SOC設置有無とアラート処理



SOC設置率



SOC設置企業は24時間以内の対応により自信があり、アラートをより適切に処理できている一方で、SOC設置率は低く留まっている

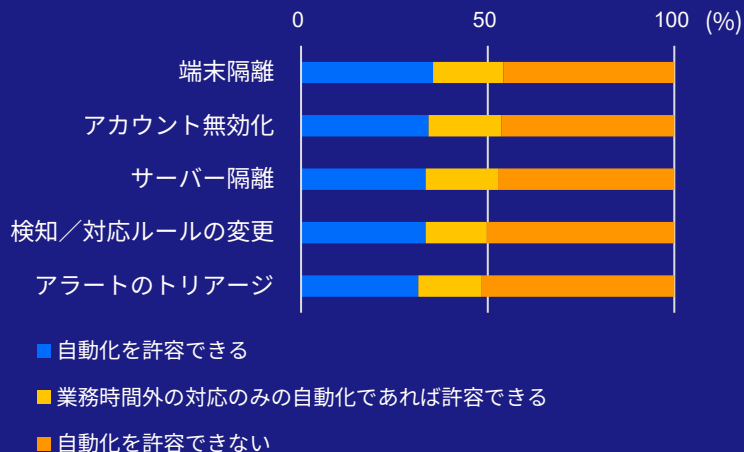


生成AI／自動化は信頼されているか？

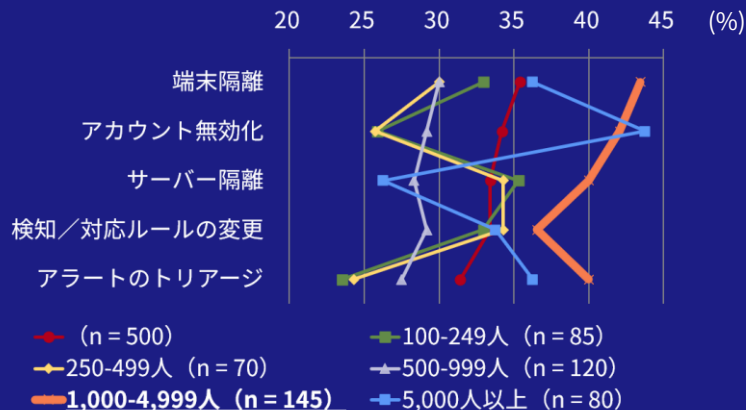


自動化の許容度はプロセスより企業規模による違いが大きい

プロセス別に見た自動化の許容度



企業規模別に見た自動化の許容度



どのプロセスにおいても、約半数の回答者が自動化を許容できないものの、従業員が1,000-4,999人の企業は他の規模の企業より自動化の許容度が高い



業務利用において、AIの入出力の両方に懸念がある

生成AIおよびAIエージェントに 関して懸念しているリスク

全体で懸念の割合が最も高い

1,000人以上の企業で
懸念の割合が最も高い

主な脆弱性パターン：

- パッケージハルシネーション
(スロップスクワッティング)
- ハードコードされた認証情報
- コードハルシネーション

外部のAIへの機密情報入力

AI生成コードの脆弱性

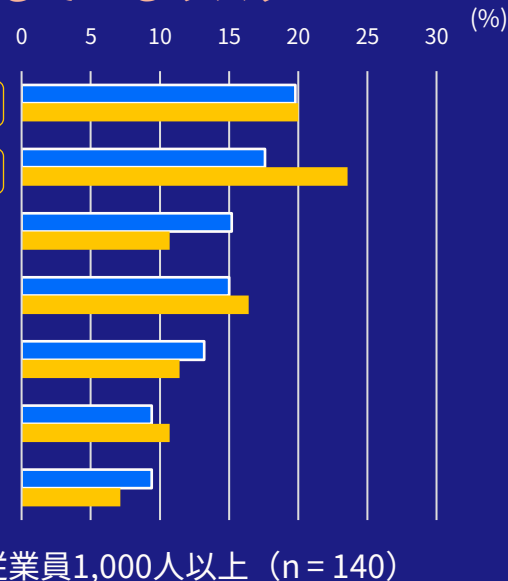
AIによる誤情報の出力

提供事業者側の情報漏洩

シャドーAIによる情報漏洩

著作権の侵害や法令違反

懸念しているものはない



AIエージェントの行動は信頼できるか？



MCPは“AIのためのUSB-C”



ID管理／アクセス制御の要素で見たMCPの課題

	Identification 識別	Authentication 認証	Authorization 認可	Accounting 記録
MCP	OAuth2.1に準拠し カバー 実装依存の識別は可能 • OAuth client ID • サーバ名 • mTLS証明書の主体名	ある程度できる • mTLSによる接続相手 認証 • OAuthトークンによる クライアントの確認	ツール単位の制御 • ClientがServerを呼 べるか • HTTP上でOAuthを 使った認可	システム観点の記録 • サーバログ • APIアクセス記録
新たな標準化 が議論されて いる項目 “KYA-OS”	Agentの識別子 • DID/Agent ID • 継続的に識別可能	行為主体としての 真正性確認 • 識別子を鍵、証明書 で認証	誰の代理で、 どこまで許可か • 親AIエージェントの 権限を超えていない か • 失効確認／ライフサ イクル管理	どのAIエージェント が何をしたか • 追跡、監査可能

セキュリティに関して実装依存の範囲が広く、新たな標準化が議論されている

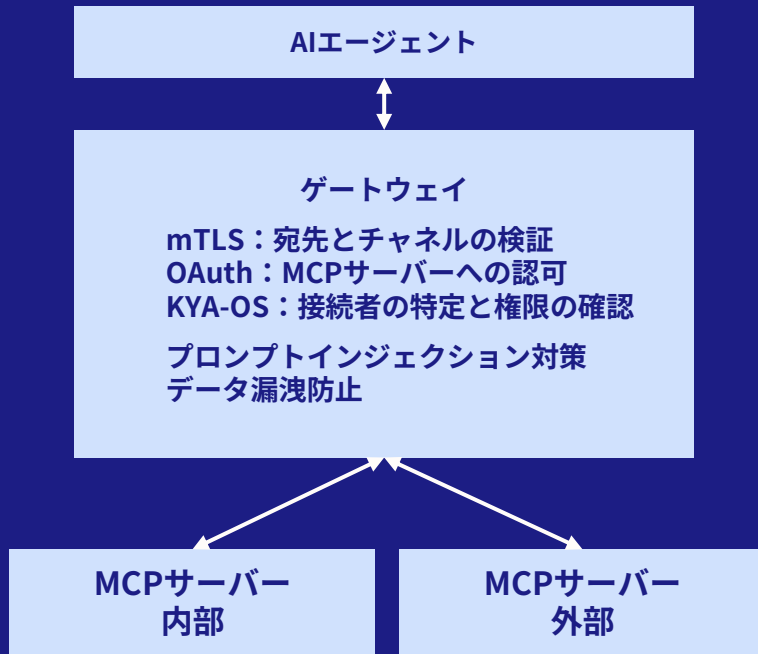


MCPプロキシ／ゲートウェイ

プロキシ：各MCPサーバーの前面でIDとパーミッションを管理



ゲートウェイ：すべてのMCP通信を一元的に可視化／制御



提言



ユーザー企業が感じる課題



AIによる自動対応の判断根拠を経営層や現場に説明できず、導入／拡大の承認が得られない



自動化機能やAIエージェントに何をどこまで任せてよいか社内基準がなく、活用の範囲を広げられない



セキュリティサービスを提供するベンダーへの提言： 信頼を製品戦略の中核にすべきである



判断根拠の可視化
を製品設計に組み
込むべきである



顧客が自社の「自動化の
許容度」を明らかにする
ためのコンサルティング
を提供すべきである



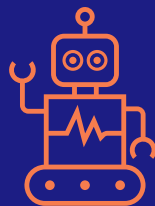
実績データや事
例を積極的に開
示すべきである



AIエージェントのための「ID管理の4要素」
の機能を製品に組み込むべきである
識別、認証、認可、記録



セキュリティサービスを提供しないベンダーへの提言



顧客企業のセキュリティ審査の要件が高まる中、セキュリティ機能を提供しなくとも、AIを活用した製品・サービスであれば、「信頼の設計」は避けて通れない



判断根拠の可視化、データ取り扱い方針の明示、エージェント動作の制御機能は、顧客企業のセキュリティ審査における評価基準となり、製品競争力の差別化要因となる

**自社製品／サービスに組み込まれたAIに
「信頼の設計」を実装すべきである**



 IDC
DIRECTIONS