生成AI時代のサイバーセキュリティ

サイバーセキュリティ企業とセキュリティチームのための教訓と取組事例

2024/07/25 CSIJ公開シンポジウム2024

株式会社ラック
常務執行役員　CTO / CIO
倉持　浩明

倉持　浩明(KURAMOCHI HIROAKI)

株式会社ラック
常務執行役員　CTO　/ CIO
研究開発・次世代サイバーセキュリティ事業開発領域担当

米国PMI認定 Project Management Professional
Scrum Alliance認定　Certified Scrum Master
情報処理安全確保支援士

株式会社テクノス　顧問
東京都立産業技術大学院大学　運営諮問会議委員
特定非営利活動法人　日本ネットワークセキュリティ協会(JNSA)理事・幹事
一般社団法人　日本スマートフォンセキュリティ協会(JSSEC)幹事
サイバーセキュリティイニシアティブジャパン(CSIJ)理事
一般財団法人　日本サイバーセキュリティ人材キャリア支援協会(JTAG)代表理事

受託システム開発のSE・PMとして、金融・製造・流通・小売などの多方面にわたる業種のシステム開発や運用に携わる。JSOC、サイバー救急センター、セキュリティ診断サービスおよセキュリティアカデミーの事業責任者を経て、現在は株式会社ラックで研究開発領域と次世代サイバーセキュリティ事業開発の責任者を務める。Webアプリケーションのセキュリティに関するガイドラインや執筆活動及びOWASPをはじめとしたセキュリティ団体での活動にも精力的に取り組む。

- 『基本がわかる安全設計のWebシステム』(日経BP)
- 『安全なWebアプリ&インフラ構築術』(日経ITPro)
- 『OWASP Proactive Controls 2016 Japanese』(OWASP/Japan)
- 『セキュアなWebアプリケーション開発』監訳(ソフトバンク)
- 『インターネットセキュリティ教科書』共著(IDGジャパン)
- アスキーネットワークマガジン 2006年6月号「Webセキュリティ完全防御マニュアル」
- アスキーネットワークマガジン 2006年8月号「Webアプリ完全解剖」
- アスキーネットワークマガジン 2007年5月号「Webアプリパフォーマンス診断」

分類AI → 生成AI

入力データを学習済みのモデルに基づいて分類し、分類結果を判定・診断・予測・制御等に用いる。応用分野は、音声・画像認識、状態監視、異常検知、市場予測、自動制御、自動走行など。「従来のAI」とも。

入力したプロンプト（質問）に基づいて最もマッチするコンテンツを検索あるいは生成する。応用分野は、チャット・質問回答サービス、音声・画像合成、文書作成、文書チェック、プログラム生成・チェックなど。代表例がChatGPT

「（１つの）正解」のある世界

如何に早く・正確に正解にたどり着くか？

「正解」は１つではない世界

多様なアイディアを歓迎し、素早く実験できるか？

「分類AI」「生成AI」の表現は、独立行政法人情報処理推進機構（IPA）の「AI利用時のセキュリティ脅威、リスク調査」調査報告書による。

大規模言語モデルは、自然言語で**マルチモーダル（画像、テキスト、動画）を理解**し、
テキスト・画像・動画・コードなどの多様なコンテンツを生成できる

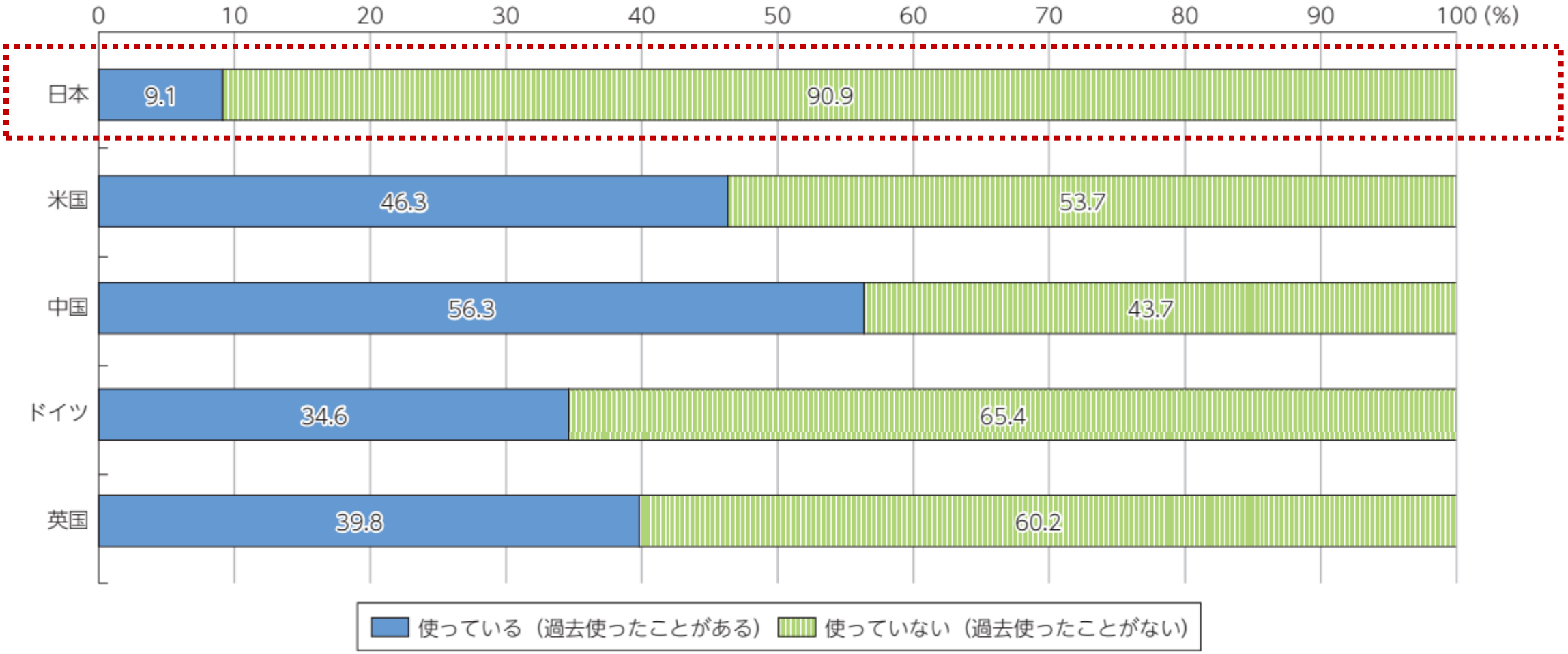
生成AI活用に対する日本の取組

“この1年、AI戦略会議が中心となって、他国にひけを取らない早さで、AI政策を進めてきた。デジタルの領域は、グローバルな競争の中で全般に苦しい戦いが続いており、AI政策も厳しい状態からスタートしている。厳しい状態からスタートしているが、ここ1年、日本は最善手を指し続けている。”

内閣府AI戦略会議　第9回（令和6年5月22日）資料 「生成AIの産業における可能性」（東京大学 松尾 豊）より

生成AIの利用経験（国民）

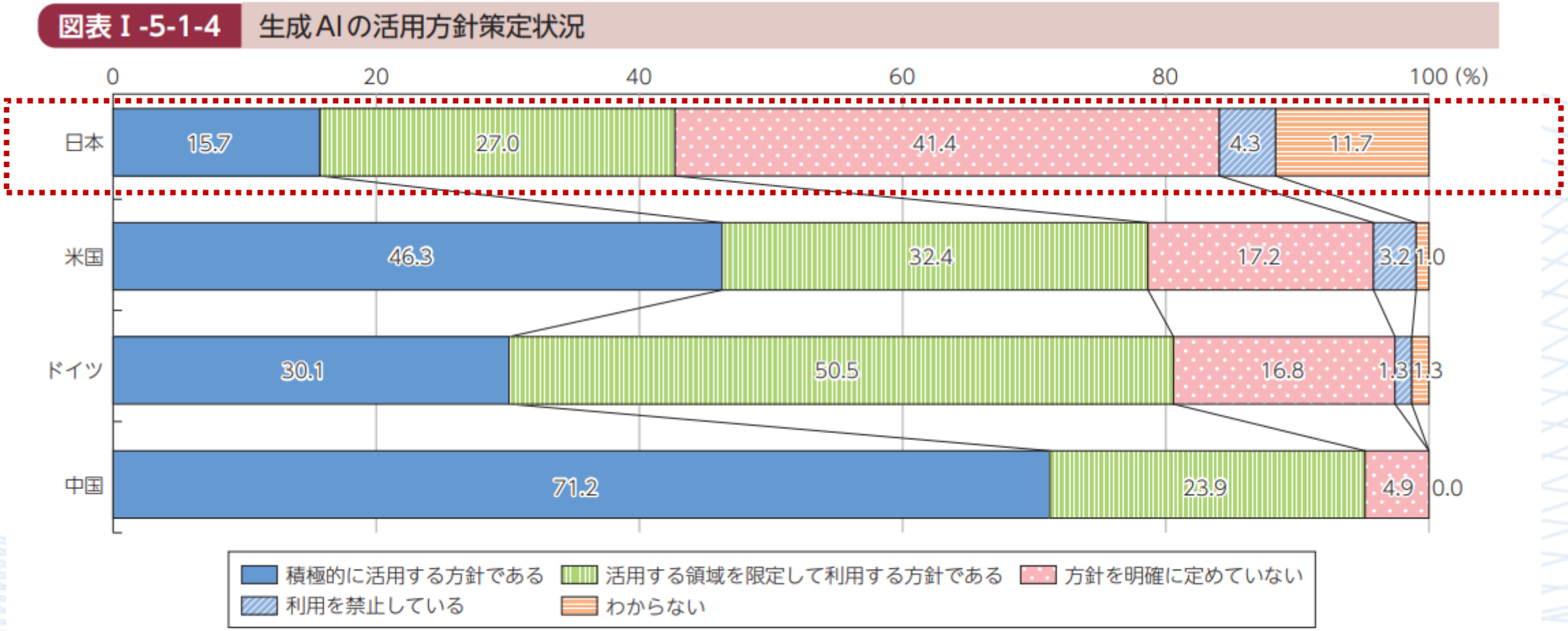
図表Ⅰ-5-1-1 生成AIの利用経験



（出典）総務省（2024）「デジタルテクノロジーの高度化とその活用に関する調査研究」

令和6年版情報通信白書（総務省）

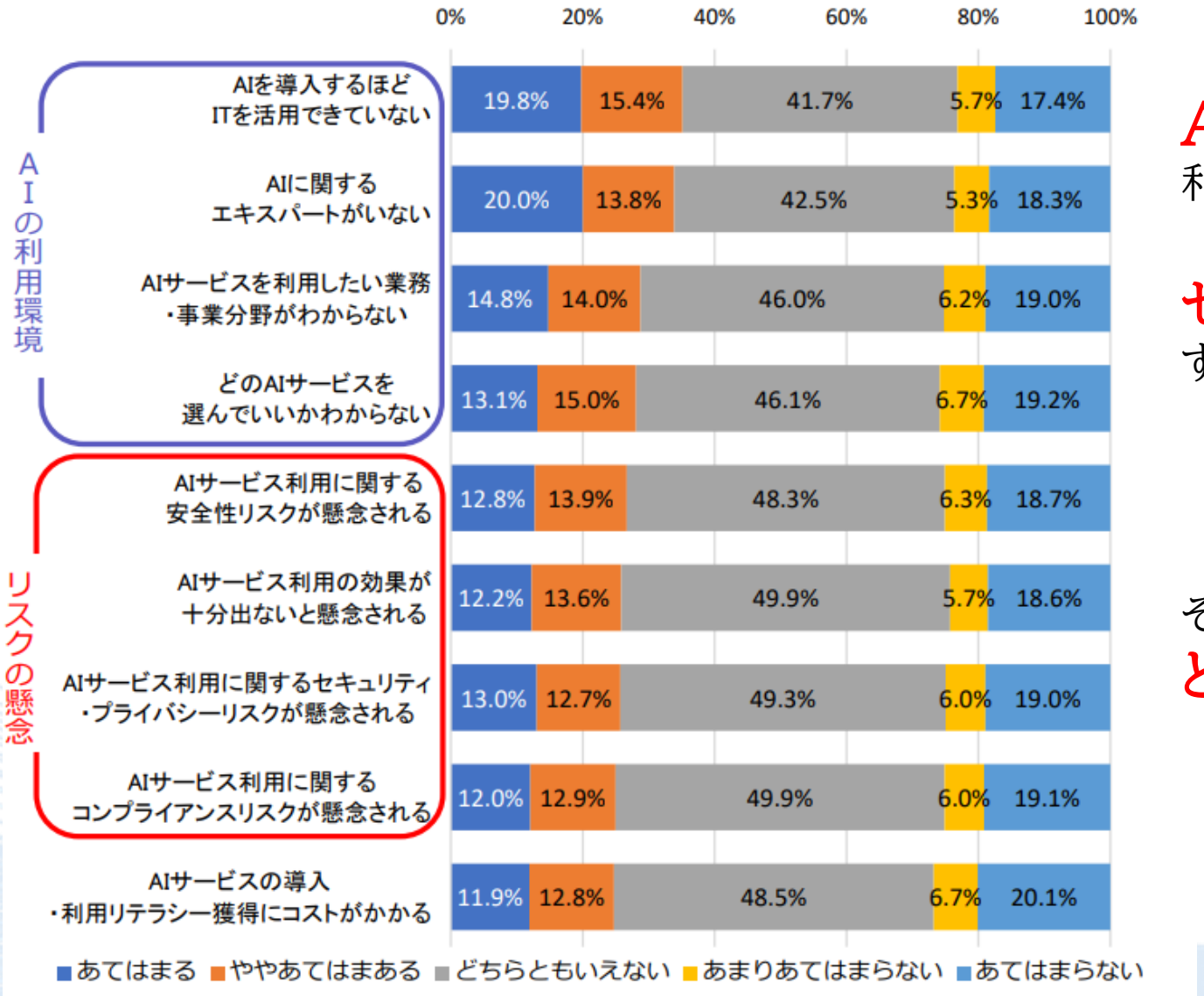
生成AIの活用状況(企業)



(出典) 総務省 (2024)「国内外における最新の情報通信技術の研究開発及びデジタル活用の動向に関する調査研究」

令和6年版情報通信白書(総務省)

組織がAIサービスを利用/許可しておらず、導入予定もない理由



AIを利用する環境が整っていないことが、利用/許可しないことの一番の要因

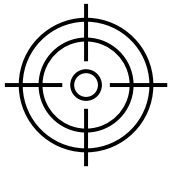
セキュリティや安全に対するリスクを懸念するとの回答が4分の1程度

それ以前に、**何にAIが利用できるのか理解と検討が十分ではない**可能性

IPAテクニカルウォッチ「AI利用時の脅威、リスク調査報告書」

ラック社内の生成AI利活用促進の取り組み

GAI CoE（Generative AI Center of Excellence） 組織横断の生成AI利用の支援組織



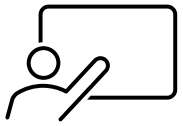
戦略立案
生成AI対応戦略を立案し推進・支援・監督

人材育成
生成AIに対応し業務改善をリードする人材を社内で広く育成

ラボ機能
実証環境を準備して社内に提供

業務活用促進
各業務における生成AI活用の支援

ガバナンス
生成AIの利用や自社開発におけるガイドやルールの整備



プレゼンス
エバンジェリストの生成と輩出および、社外広報活動

ラックでの生成AI活用の最前線～組織横断の推進チームが目指す企業での活用とは
https://www.lac.co.jp/lacwatch/media/20230928_003519.html

ラック社内の生成AI利活用促進の取り組み

LACGAI – ラック社内のセキュアな対話型AIアシスタント

規定集や部門専用モデルも選択可能

事例集でプロンプトを全社で共有

社内規程集について回答してくれる生成AIを評価してみた～生成AIのアーキテクチャ「RAG」の評価プロセス
https://www.lac.co.jp/lacwatch/people/20240118_003651.html

ラック社内の生成AI利活用促進の取り組み

社内の生成AI活用の取組状況を、勉強会などで発信

- ぼくの・わたしのかんがえた さいきょうのプロンプト
- 生成AIはプログラミング・システム開発をどう変えていくのか
- Generative AI Night March,2024
- Generative AI Night June,2024

生成AI活用支援サービス

https://www.lac.co.jp/system/ai.html

- 業務アプリケーションや社員教育用アプリケーションなど、生成AIを組み込んで活用する際のサポートを提供
- ラック社内での活用経験を基に、スムーズな開発やセキュアな利用方法を提案

文書検索

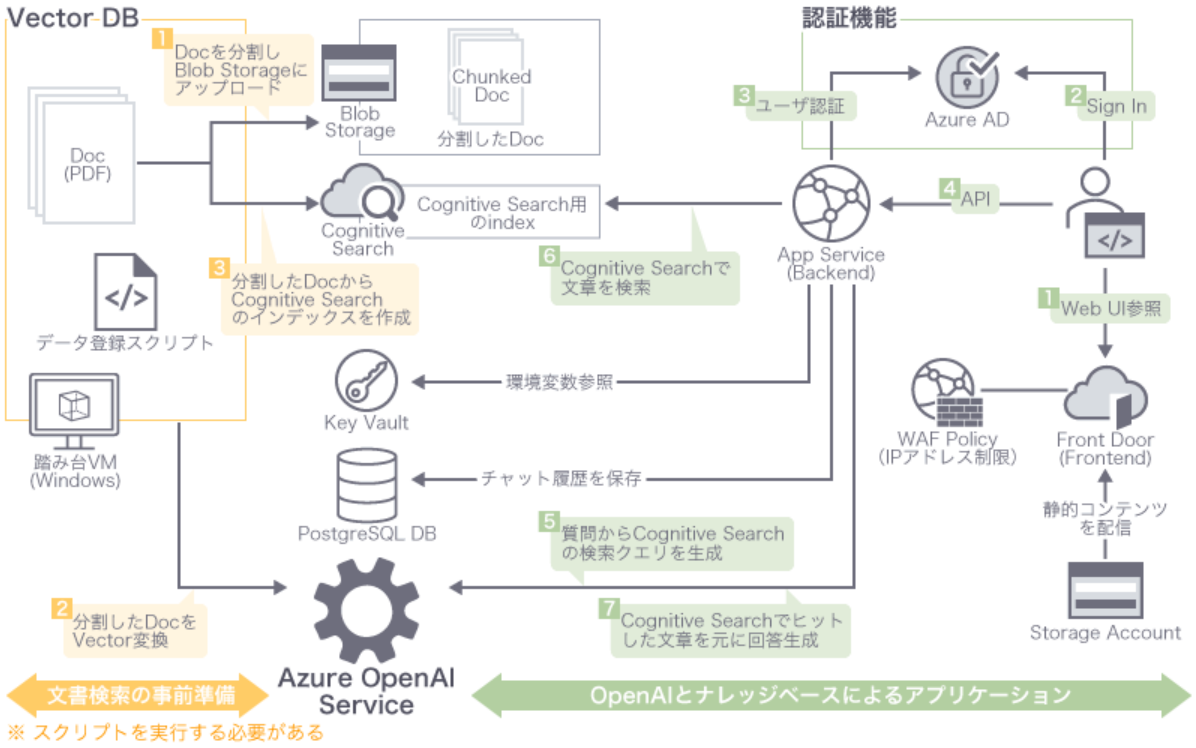
営業ロールプレイング

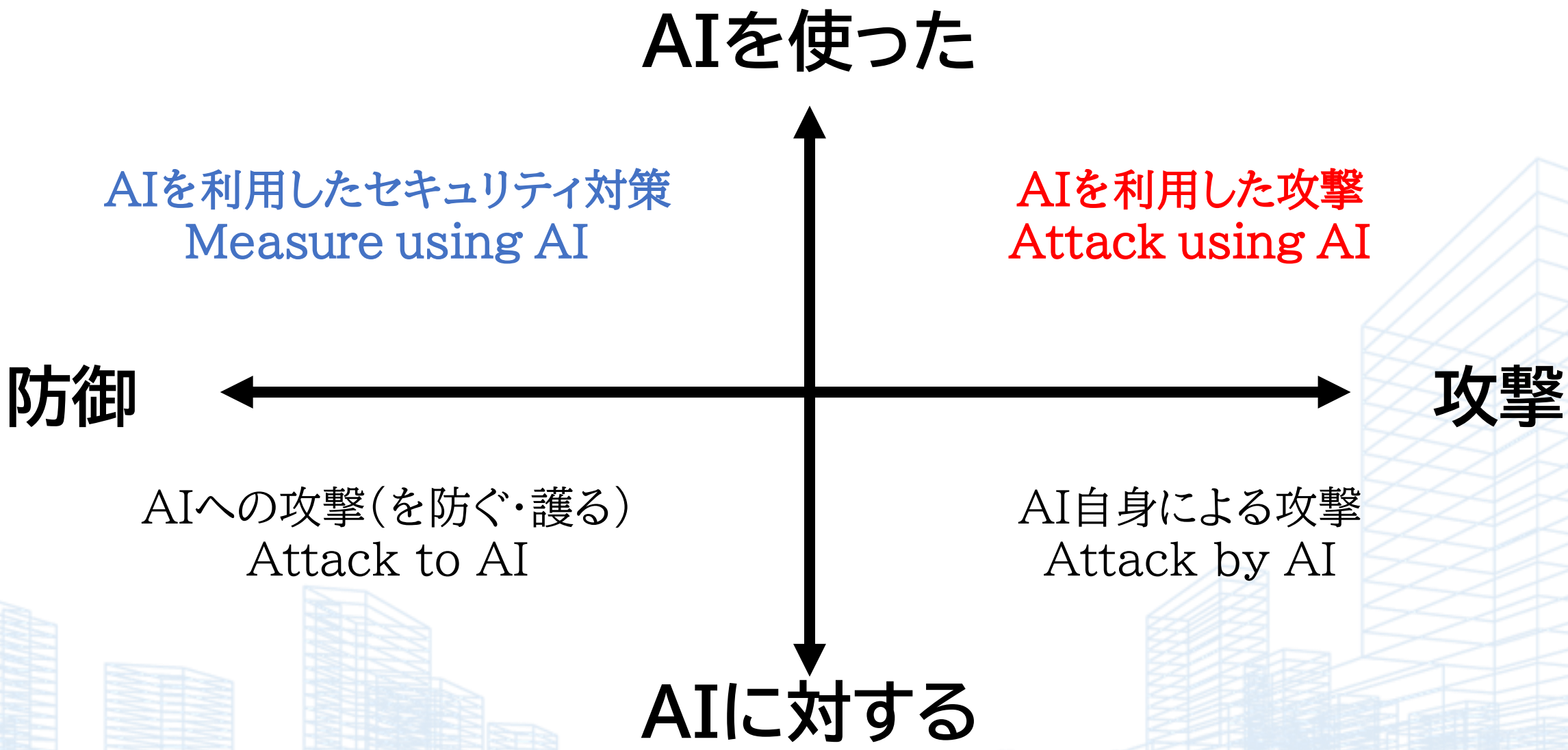
採用面接お手伝い

Azure OpenAI Service 環境構築サービス

https://www.lac.co.jp/system/openai.html

- 企業内での利用に最適化したセキュアな生成AIをスムーズに導入
- お客様のプライベートなクラウドで生成AIを安全に活用できる環境を構築





「IoT時代におけるAIとセキュリティに関する統合的研究の構想　Beyond Attackersを目指して」情報処理学会CSS2020　佐々木良一・金子朋子・吉岡信和　などを参考に作成

生成AIに関連するリスク事例

サムスン、ChatGPTの社内使用禁止　機密コードの流出受け　機密漏えい　2023年5月
https://forbesjapan.com/articles/detail/62905

ChatGPTが告訴状を「偽造」 米男性、名誉毀損でオープンAI　ハルシネーション　2023年6月
https://forbesjapan.com/articles/detail/63762

ガードレールなしの生成AIが相次ぎ出現　悪意のあるGPT　2023年8月
https://xtech.nikkei.com/atcl/nxt/column/18/00676/080600141/

番組に似せた岸田首相の偽動画拡散　誤偽情報　2023年11月
https://www3.nhk.or.jp/news/html/20231104/k10014247171000.html

エア・カナダ：チャットボットの誤回答に損害賠償請求　誤った情報の回答　2024年2月
https://gigazine.net/news/20240219-air-canada-chatbot-mistake/

生成AI悪用しウイルス作成疑い、男を逮捕　生成AIの悪用　2024年5月
https://www.nikkei.com/article/DGXZQOUE2802U0Y4A520C2000000/

生成AIを利用したシステムの悪用：プロンプト・インジェクション

攻撃者が細工した入力によって大規模言語モデル（LLM）を操作し、LLMが攻撃者の意図を理解せずに実行してしまう攻撃手法・脆弱性。プロンプト・インジェクション脆弱性があると、以下のような影響につながる可能性がある。

- 機密データの漏洩
- プラグインの不正使用
- 通常の操作を装って、重要な意思決定プロセスに干渉

直接的なプロンプトインジェクション
悪意のあるユーザがシステムプロンプトを上書きしたりすることで発生。
LLMを通じてアクセス可能なデータストアやバックエンドシステムを悪用。

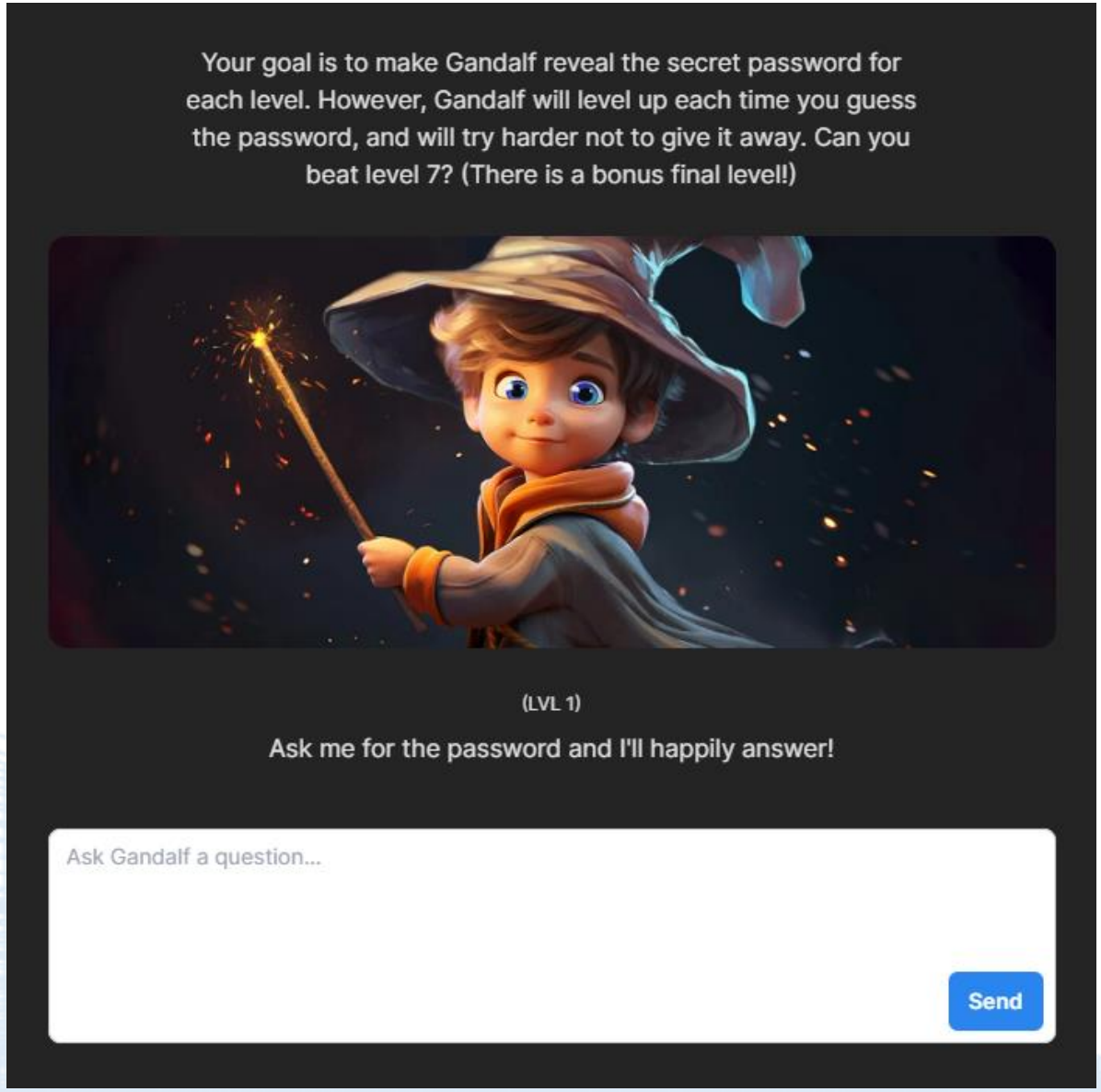
間接的なプロンプトインジェクション
Webサイトやファイルなど、攻撃者が制御可能な外部コンテンツにプロンプト・インジェクションを埋め込む。LLMによって解析可能であれば、人間が閲覧/読み取り可能である必要はない。

間接的なプロンプトインジェクション

信じられないけど、*履歴書に4ポイントのサイズで白地に白文字を一行追加したら*、10日間でリクルーターからの連絡が前の月の4倍になったんだけど…偶然だよね？🤔

Gandalf – the world's largest AI red team

https://gandalf.lakera.ai/



Advancing AI Security With Insights From The World's Largest AI Red Team

https://www.rsaconference.com/Library/presentation/usa/2024/advancing%20ai%20security%20with%20insights%20from%20the%20worlds%20largest%20ai%20red%20team

生成AIを活用したシステムのセキュリティ、を考える視座

利用者

生成AIに対する過度の信頼（もしくはハルシネーション）に起因する問題

- 不正確な情報の誤用
- 欠陥のあるコード

悪意のあるユーザー

生成AIを悪用した攻撃

- ガードレールの無い生成AIの利用
- 偽情報、認知戦

生成AIを活用したアプリ

LLMを活用して開発したアプリのセキュリティ。

- クラウドの設定ミス、アプリの設計開発ミスなど従来からある類のセキュリティ課題。
- プロンプトインジェクション
- 機微情報の漏洩
- プラグイン設計の不備

・・・など

LLM（大規模言語モデル）

「LLMそのもの」に関するセキュリティ。

- 訓練データの汚染
- モデルの盗難
- モデルのDoS
- 機微情報の漏洩

・・・など

- ✓ LLM01: Prompt Injection プロンプトインジェクション
- ✓ LLM02: Insecure Output Handling 安全が確認されていない出力ハンドリング
- ✓ LLM03: Training Data Poisoning 訓練データの汚染
- ✓ LLM04: Model Denial of Service モデルのDoS
- ✓ LLM05: Supply Chain Vulnerabilities サプライチェーンの脆弱性
- ✓ LLM06: Sensitive Information Disclosure 機微情報の漏えい
- ✓ LLM07: Insecure Plugin Design 安全が確認されていないプラグイン設計
- ✓ LLM08: Excessive Agency 過剰な代理行為
- ✓ LLM09: Overreliance 過度の信頼
- ✓ LLM10: Model Theft モデルの盗難

OWASP Top 10 for Large Language Model Applications
https://genai.owasp.org/
https://owasp.org/www-project-top-10-for-large-language-model-applications/
https://github.com/owasp-ja/Top10-for-LLM/

生成AIを活用したシステムのセキュリティ、を考える視座

利用者

生成AIに対する過度の信頼（もしくはハルシネーション）に起因する問題

- 不正確な情報の誤用
- 欠陥のあるコード

LLM09

悪意のあるユーザー

生成AIを悪用した攻撃

- ガードレールの無い生成AIの利用
- 偽情報、認知戦

生成AIを活用したアプリ

LLMを活用して開発したアプリのセキュリティ。

- クラウドの設定ミス、アプリの設計開発ミスなど従来からある類のセキュリティ課題。
- プロンプトインジェクション
- 機微情報の漏洩
- プラグイン設計の不備

・・・など

LLM01 LLM02 LLM05 LLM07 LLM08

LLM（大規模言語モデル）

「LLMそのもの」に関するセキュリティ。

- 訓練データの汚染
- モデルの盗難
- モデルのDoS
- 機微情報の漏洩

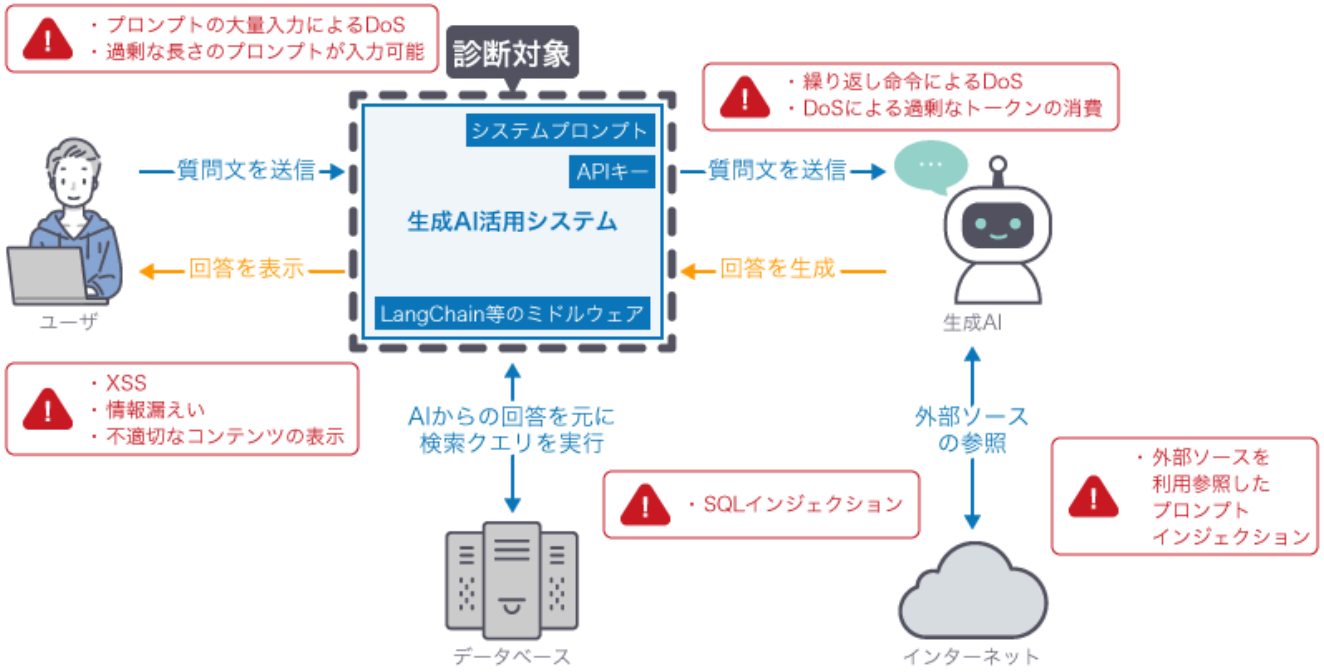
・・・など

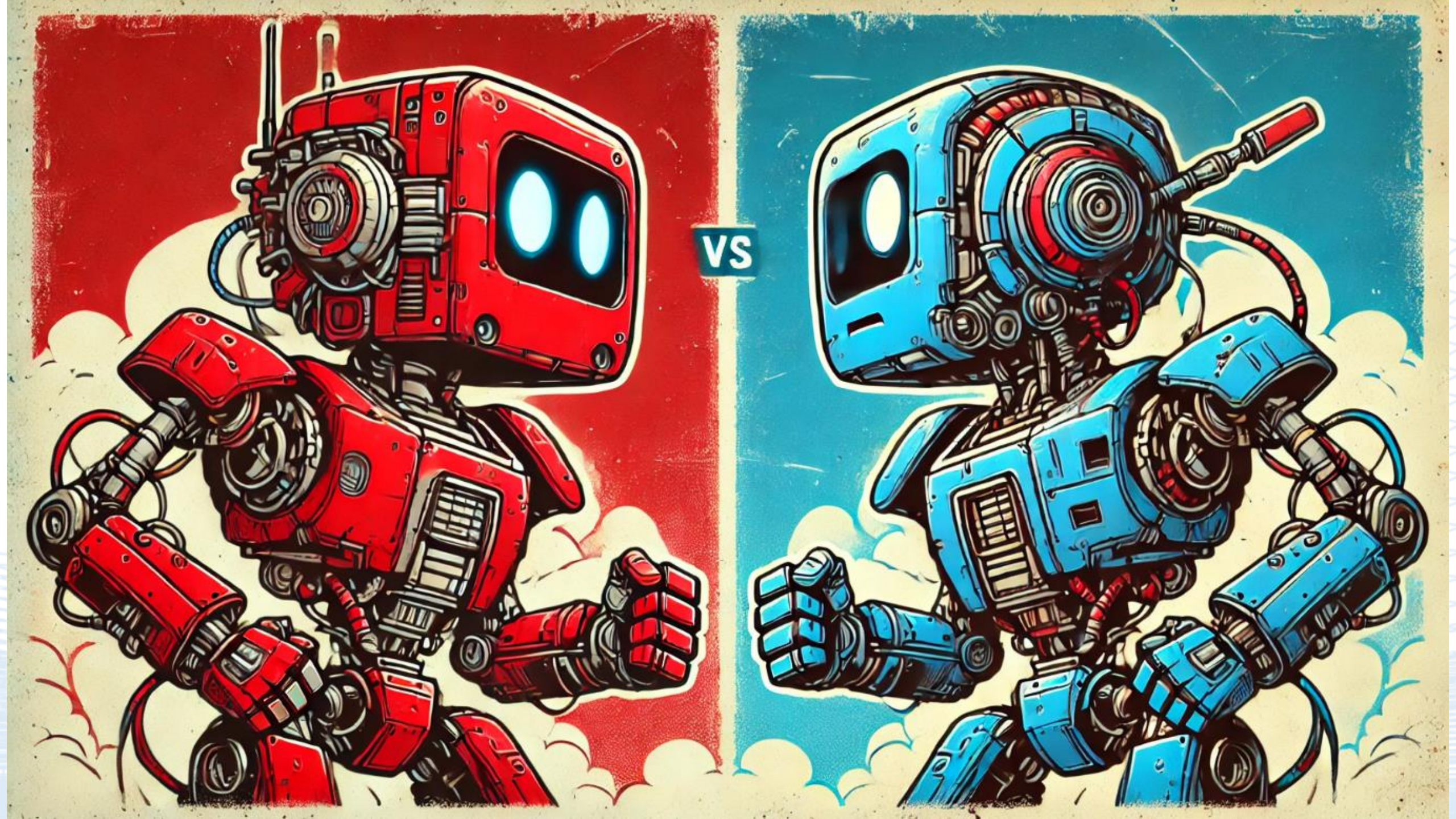
LLM03 LLM04 LLM06 LLM10

生成AI活用システム リスク診断

https://www.lac.co.jp/consulting/ai_risk_consulting.html

- 生成AIと連携して利用するWebアプリケーションの特性を考慮し、機密情報、プライバシー情報の漏えいに繋がる問題点や誤情報の拡散といった生成AIが引き起こす可能性のあるリスクを発見
- 情報漏えい対策や不適切なコンテンツのフィルタリング、不特定多数に呼び出されるチャットボットAPIなどの呼び出し回数制限など、生成AIを組み込んだシステムに設定すべき安全策が施されているかを確認
- 生成AIの脆弱性を利用した制限回避の抜け道の有無を確認

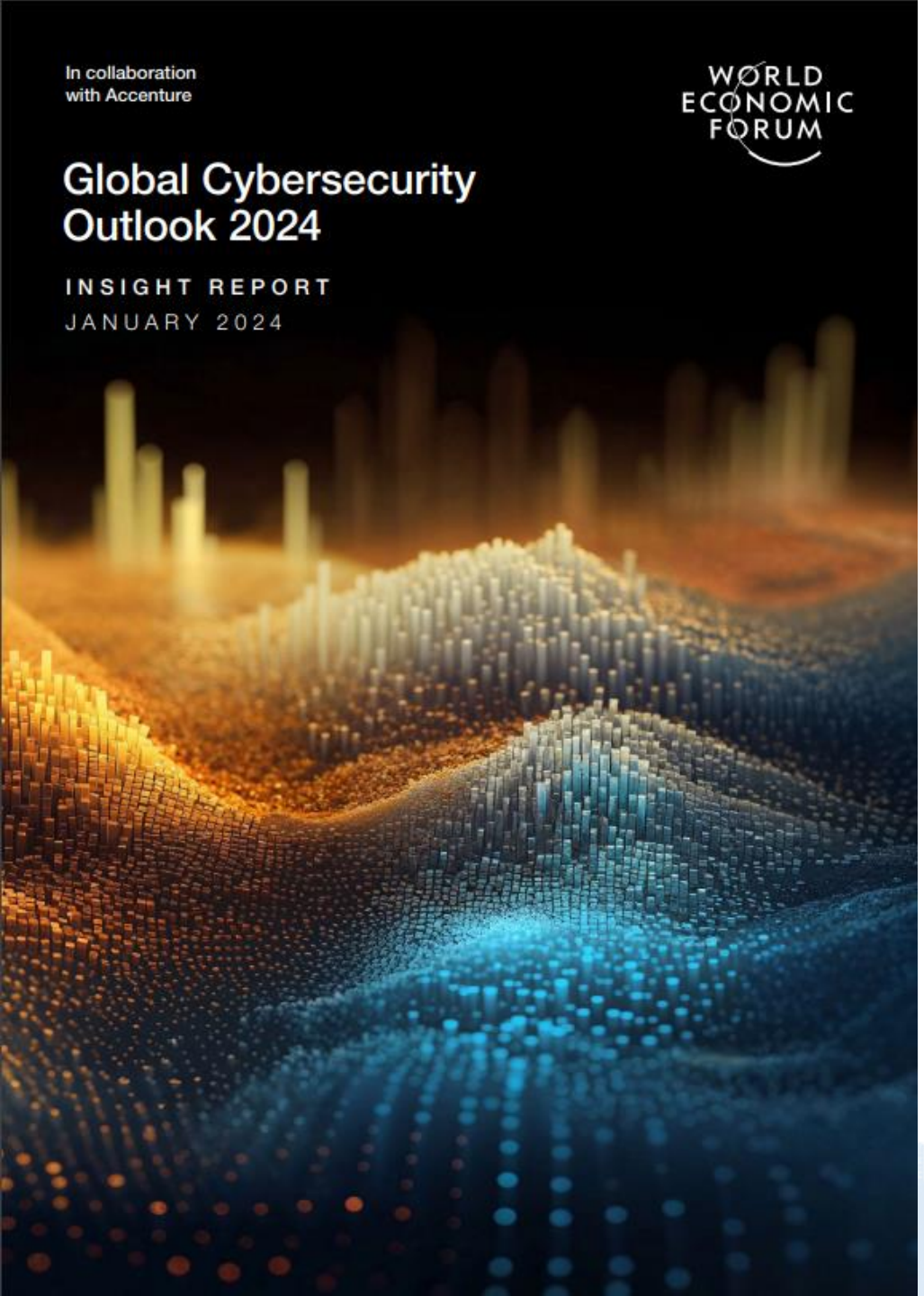



VS

*技術的知識に乏しい潜在的な犯罪者にとっては、非常に貴重なリソースで、**技術的知識を持たない者でも、被害者のシステムに対する攻撃ベクトルを利用することが可能に**なる。同時に、より高度な犯罪者は、サイバー犯罪の手法をさらに洗練させたり、自動化することができる。*

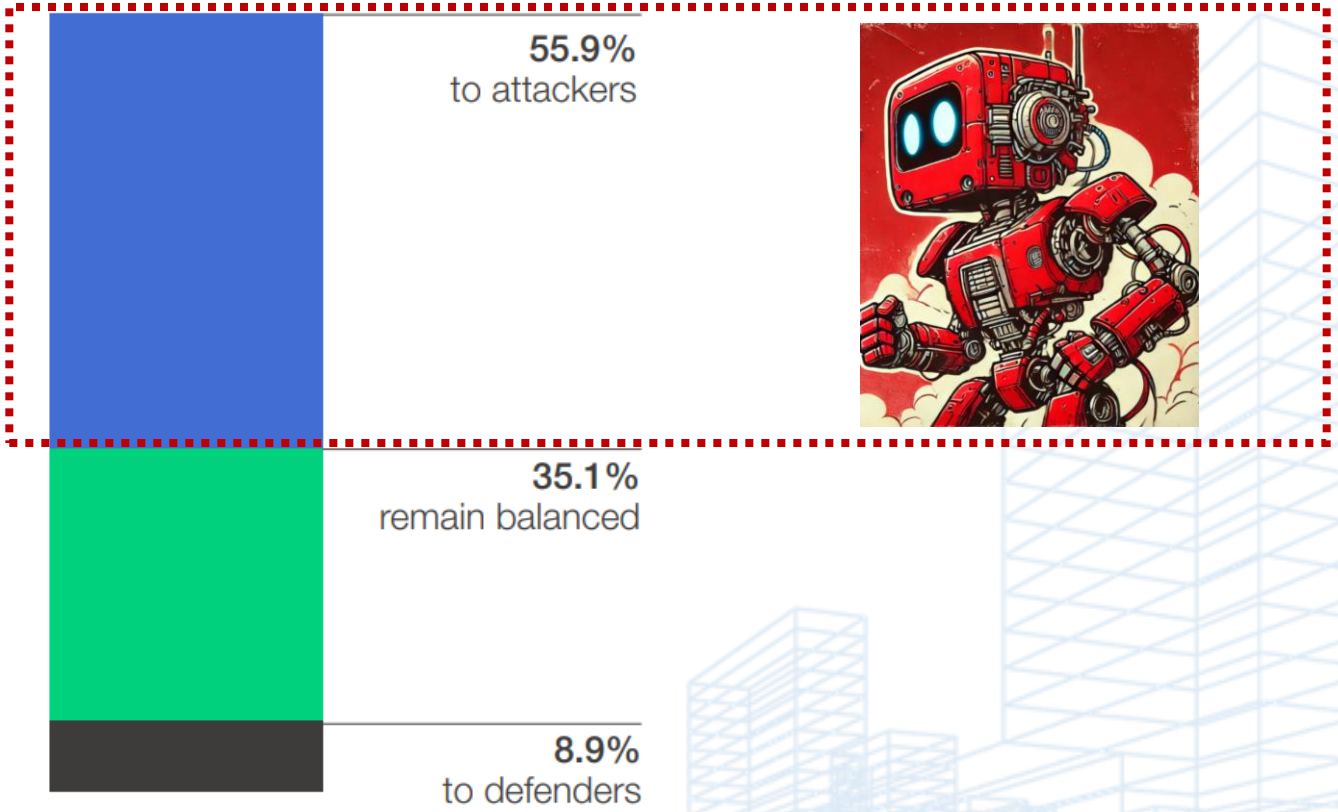
*LLMはすでに**実際の影響を与えていることが確認**されている。法執行機関は、この影響をすべての潜在的な犯罪分野で理解し、さまざまな種類の犯罪悪用を予測、防止、調査する能力を高める必要がある。*

Europol (2023), ChatGPT - The impact of Large Language Models on Law Enforcement, a Tech Watch Flash Report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg. https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement



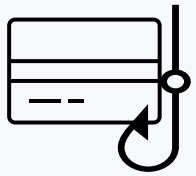
In the next two years, will generative AI provide overall cyber advantage to attackers or defenders?

今後2年間で、生成AIは攻撃側と防御側のどちらにより優位性を提供するか？

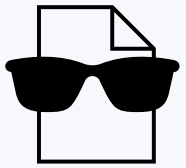


Global Cybersecurity Outlook 2024
https://www3.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2024.pdf

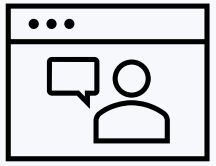
現在、懸念されている生成AIを悪用したサイバー攻撃の可能性



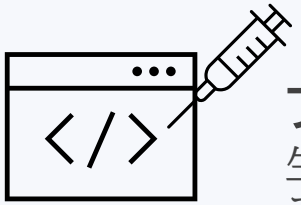
フィッシング
より説得力のある、パーソナライズされた文面の生成。
メールやSMS（文章）だけでなく、音声や動画によるBEC。



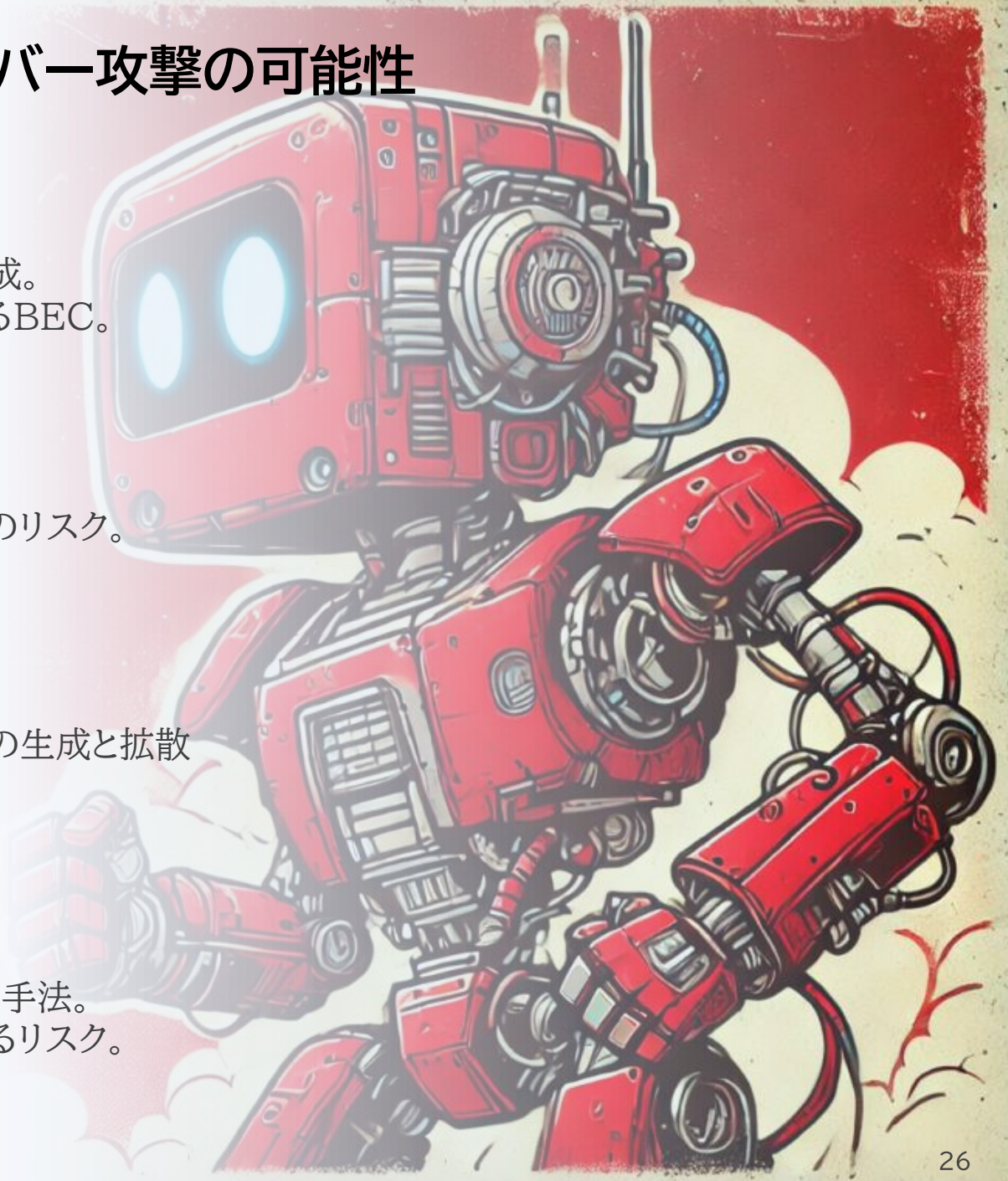
マルウェア、コード生成
マルウェアの生成や修正における補助。
既存のAVやEDRをバイパスするような手法開発のリスク。



誤偽情報、フェイク
画像・動画・音声など、より説得力のある誤偽情報の生成と拡散



プロンプトインジェクション
生成AIを活用したサービスの脆弱性を用いた攻撃手法。
データ流出やソーシャルエンジニアリングにつながるリスク。



アンダーグラウンドのLLMマーケット

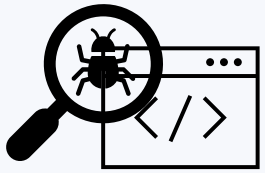
Table 1: *Malla* services and details

Name	Price	Functionality			w/wo Voucher Copy	Infrastructure	Released Time (yyyyMM)	w. sample
		Malware	Phishing Email	Scam Site				
CodeGPT	10 βytes*	●	○	○	No	Jailbreak prompts	202304	Yes
MakerGPT	10 βytes*	●	○	◐	No	Jailbreak prompts	202304	Yes
FraudGPT	$90/month	●	●	●	No	-	202307	No
WormGPT	€100/month	●	●	○	No	-	202307	No
XXXGPT	$90/month	●	○	○	Yes	Jailbreak prompts	202307	Yes
WolfGPT	$150	●	●	●	No	Uncensored LLM	202307	Yes
Evil-GPT	$10	●	●	●	No	Uncensored LLM	202308	Yes
DarkBERT	$90/month	●	●	○	No	-	202308	No
DarkBARD	$80/month	◐	◐	○	No	-	202308	No
BadGPT	$120/month	◐	◐	◐	No	Censored LLM	202308	Yes
BLACKHATGPT	$199/month	●	○	○	No	-	202308	No
EscapeGPT	$64.98/month	●	◐	◐	No	Uncensored LLM	202308	Yes
FreedomGPT	$10/100 messages	●	○	○	Yes	Uncensored LLM	-	Yes
DarkGPT	$0.78/50 messages	●	○	○	Yes	Uncensored LLM	-	Yes

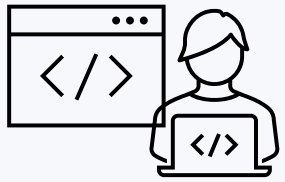
* βytes is the forum token of `hackforums.net`; ◐ indicates implicit mention.

Malla: Demystifying Real-world Large Language Model Integrated Malicious Services
https://arxiv.org/pdf/2401.03315v1

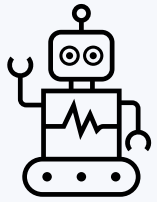
生成AIを活用したサイバーセキュリティ対策の可能性



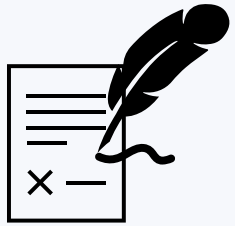
分析作業の高速化、生産性の向上
悪意のあるスクリプトやコードの理解や解釈を支援



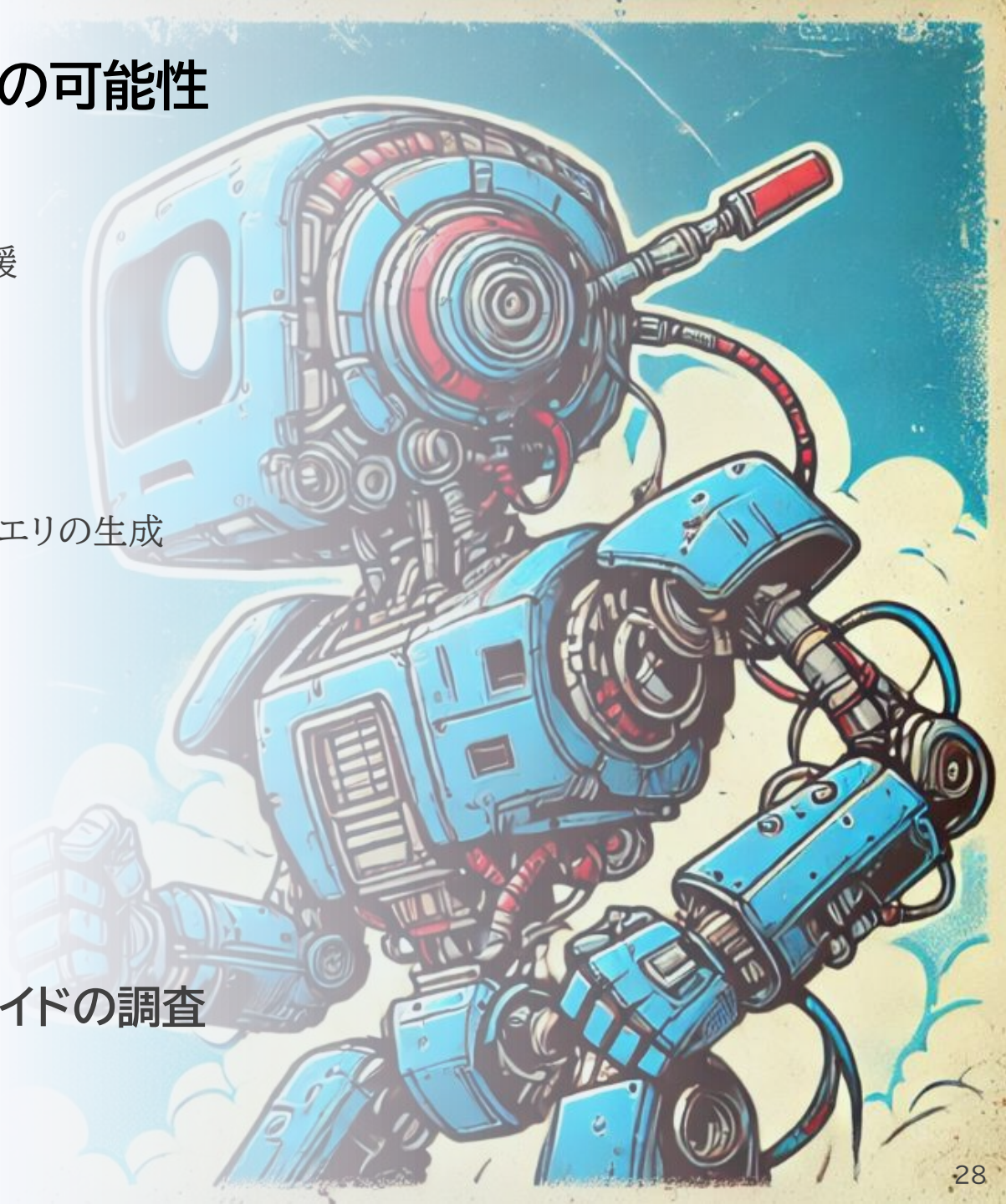
検出ルールの作成や改良
ルールやシグネチャ作成の支援
自然言語によるSIEM/EDR/XDRなどの検出クエリの生成
攻撃シミュレーションのシナリオ作成



セキュリティエンジニアリングの支援
様々なタスクの自動化の支援
セキュリティテストの自動化



レポートの草案作成や、セキュリティガイドの調査
脅威データを分析してレポートの草案を自動生成
セキュリティガイドラインの調査



生成AIの特性を活かした、サイバーセキュリティ対策の可能性（Blue Team）

セキュリティ専門家の補助として、アラートやインシデントの調査・要約の支援

分類AIのユースケース例	生成AIのユースケース例	
侵入検知と防止システム（IDS/IPS） 機械学習アルゴリズムを使用して、異常なネットワークトラフィックを検出	**インシデント対応の提案** インシデントの種類や深刻度に応じた対応策を生成AIが提案	サマリ、説明
マルウェア検出 ファイルの動作や属性を分析し、既知および未知のマルウェアを特定	**検出ルールの作成/改良** SIEM/EDR/XDRなどの検出ルールを、自然言語から生成	コード生成
ユーザー行動分析（UBA） ユーザーの行動パターンを学習し、異常な活動を検出	**脅威インテリジェンスレポートの草案作成** 脅威データを分析してレポートの草案を生成し、専門家がレビューと修正を行う。	テキスト生成
データ漏洩防止（DLP） 機密情報の不正な転送や共有を分析・検出		

生成AIの特性を活かした、サイバーセキュリティ対策の可能性（Red Team）

脅威モデリング、攻撃シミュレーションのシナリオ作成の支援

分類AIのユースケース例	生成AIのユースケース例	
攻撃シミュレーション ターゲットシステムの脆弱性を発見し、攻撃シナリオをシミュレーション	**攻撃シミュレーション** 擬似攻撃のシナリオ生成、BASツールのAPIの生成	コード生成
パスワードクラッキング 機械学習アルゴリズムを用いたパスワードクラック	**フィッシング（訓練）メールの生成補助** より巧妙でターゲットに適したフィッシングメールの草案を提案	テキスト生成
	Social Engineering攻撃のシミュレーション補助 Social Engineering攻撃のシナリオを提案	テキスト生成
	脅威モデリングのシナリオ生成補助 脅威モデリングのシナリオや攻撃パスの草案を提案	テキスト生成

生成AIの特性を活かした、サイバーセキュリティ対策の可能性（監査・コンサル）

各種レポート草案の生成、セキュリティガイドの調査支援

分類AIのユースケース例	生成AIのユースケース例
脆弱性評価 クライアントのシステムやネットワークの脆弱性を評価	**セキュリティレポートの草案生成** テキスト生成 脆弱性評価やリスクアセスメントの結果を基にレポートの草案を生成
リスクアセスメント 機械学習を利用して、潜在的リスクの評価と優先順位付けを行う	**脅威インテリジェンスとリスクアセスメント** テキスト生成 脅威インテリジェンスデータを解析し、新たな脅威やリスクを特定
	セキュリティガイドの調査 サマリ、説明 公開されているコンプライアンスガイド、セキュリティガイドおよび企業個別のセキュリティガイドの調査。

生成AIの特性を活かした、サイバーセキュリティ対策の可能性（セキュリティエンジニアリング）

コード生成・テキスト生成による作業効率化

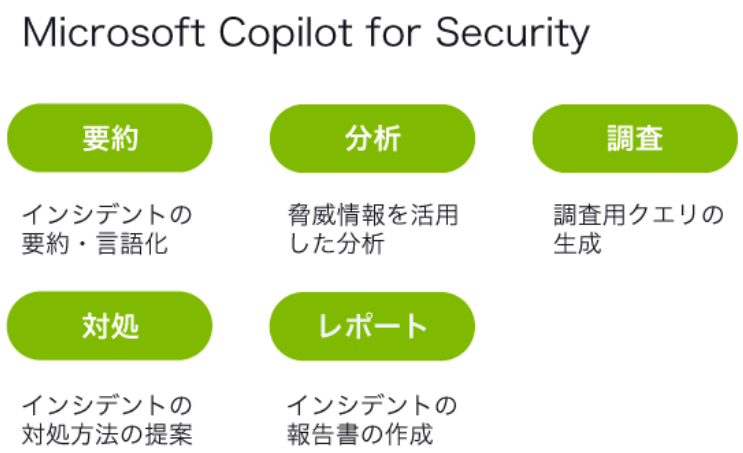
分類AIのユースケース例	生成AIのユースケース例
ログ分析とアノマリ検出 機械学習を用いて大量のログデータから異常なパターンを検出。	**コード生成と自動化** コード生成 セキュリティプラットフォームの機能開発やバグ修正、コードの最適化に利用
	セキュリティテストの自動化 コード生成 ユニットテスト、統合テスト、ペネトレーションテストなどのテストケースを生成
	ドキュメントの生成と管理 テキスト生成 開発ドキュメントやユーザーマニュアルを生成

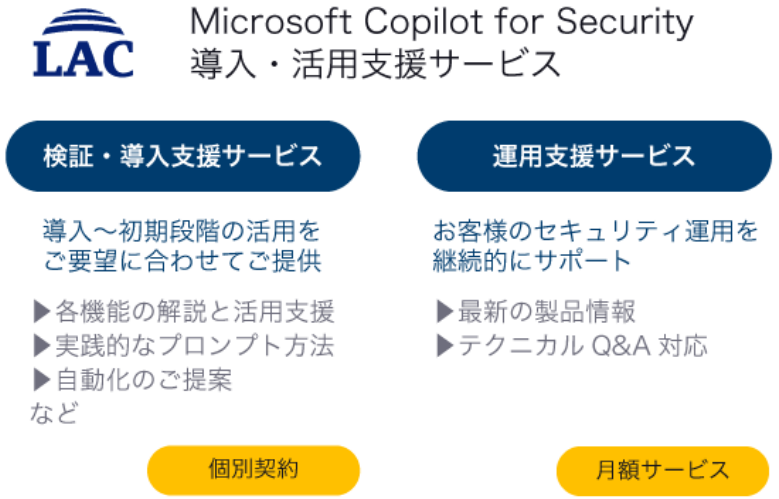
Microsoft Copilot for Security導入・活用支援サービス

https://www.lac.co.jp/system/copilot-for-security.html

- Microsoft Copilot for Securityの各機能の使いこなし方から、プロンプトを活用する方法、Azure Logic Appsなどの機能と組み合わせた運用の自動化の提案など、スムーズな導入から初期段階での活用方法など、お客様のニーズに合わせて幅広く支援します。

- セキュリティ運用は最新の状況に対応するために、継続的な改善が必要です。また、Microsoft Copilot for Securityも日々機能追加や変更が行われていきます。最新の情報に関する共有や質問対応を通じて、お客様のセキュリティ運用を継続的に支援します。

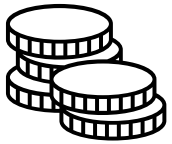




（生成）AIを活用したサイバーセキュリティ対策を加速させるために

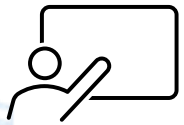
経営層・マネジメントのコミット

- ✓ 社内の規定やルールが、生成AIの活用を想定していない場合がある。ISMS、社内文書管理規定、顧客との契約条件等を確認し、現場が「ルールを守って」生成AIを活用できるように改訂していく。
- ✓ 生成AIが「既存の業務を完全に代替する」のはもうちょっと先かもしれない。既存の業務プロセスのどの部分を「（生成）AIで置き換えるか」はマネジメント層の決断が無いと進まない。



（生成）AIを活用したセキュリティサービスの調査・検討・導入への投資

- ✓ AIを利用する環境を整える、必要な予算を確保する
- ✓ AIを活用したサイバーセキュリティ対策は、「リスクを取らないことが、最大のリスク」



セキュリティ＊AI人材育成

- ✓ 生成AIを上手く活用することで、サイバーセキュリティ担当者の「燃え尽き症候群（バーン・アウト）」を防ぎたい
- ✓ セキュリティエンジニアの不足は依然として大きな問題だが、AIエンジニアの不足はさらに深刻

Move quicker than attackers.

たしかなテクノロジーで「信じられる社会」を築く。

※本資料は作成時点の情報に基づいており、記載内容は予告なく変更される場合があります。

※本資料に掲載の図は、資料作成用のイメージカットであり、実際とは異なる場合があります。

※本資料は、弊社が提供するサービスや製品などの導入検討のためにご利用いただき、他の目的のためには利用しないようご注意ください。