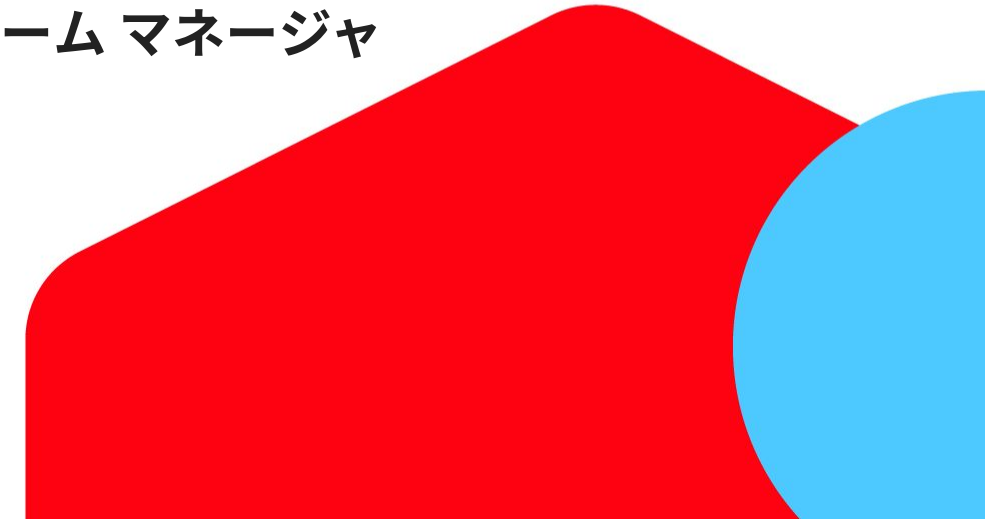
2024年2月1日 13:30pm – 15pm
デジタル時代の人材政策に関する検討会 発表資料

資料３

生成 AI 時代における
サイバーセキュリティ人材について

竹井 悠人 / Yuto Takei
株式会社メルカリ セキュリティ チーム マネージャ

情報処理技術者試験委員



mercari　mercoin

たけい　ゆうと
竹井 悠人

株式会社メルカリ
メルコイン セキュリティ チーム マネージャ

東京大学・同大学院 出身。2013年 未踏事業に採択された後、複数のスタートアップ事業を行う。2019年に株式会社 Basset を起業し、暗号資産の取引分析アルゴリズムの開発を行う。2021年、買収により株式会社メルカリに入社、メルコイン事業のセキュリティ業務に従事。同年より情報処理技術者試験委員。著書に「徹底解説 NFTの理論と実践」等。

モダンな IT 組織におけるサイバーセキュリティ業務

生成 AI の登場以前から、デジタル技術を活用する組織のセキュリティ業務は多岐に亘る

管理的

- セキュリティ施策の立案、推進。業界動向に関する情報収集
- 情報資産の管理、保有情報の棚卸し、アクセス権限の付与・剥奪
- 外部委託先のセキュリティ実施状況の評価・改善、共有情報の管理
- 関連法令・基準の把握、遵守状況の評価、適合のための対応
- 役職員向けのセキュリティ研修・教育
- セキュリティ インシデント発生時の対応統括、関係企業・当局との連携

→ DX推進スキル標準における「**サイバーセキュリティマネージャー**」のロールに相当

技術的

- 各種機器 (PC・サーバー等)、システム、ネットワーク、クラウド インフラ等における設計、構築、運用、移行、廃止等
- システムの適切な利用 (機能性・権限・コスト等) の監視
- ソフトウェア開発・セキュリティ確保のための技術的知見の活用・共有
- セキュリティ インシデント発生時の証拠保全・分析・復旧対応等

→ 同「**サイバーセキュリティエンジニア**」に相当

サイバー セキュリティに関連する業務分担の例

例えばメルカリ グループでは、大まかに下記のような業務領域 (抜粋) ごとに部門を分けてサイバー セキュリティに全社的に取り組んでいる

戦略・管理	● セキュリティ関連の情報収集、社内への教育・啓蒙 ● グループ全体的なセキュリティ戦略の策定、優先度管理
製品向けセキュリティ	● お客さま向けのアプリ、Web 等の安全性確保と脆弱性管理 ● システム全体 (インフラ・API 等) の保護、攻撃の予防と検知
社内システムセキュリティ	● 従業員の利用する IT 環境の整備と維持、権限管理 ● 脅威情報・攻撃動向の収集、ログ分析、不正アクセス検知
プライバシ対応	● 個人情報保護に関する法令等対応・施策推進 ● サービス設計上の社内向けアドバイザリ、研修の実施
各業態固有の対応	● 資金決済法・PCI-DSS・ISMS 等の個別要件に特化した対応 ● サービスの新規設計・改良時における固有のリスク管理

生成 AI 時代のサイバーセキュリティ担当者の新たな業務

生成 AI 技術の普及により、新たな状況でのセキュリティ業務が加わる可能性がある

本資料の焦点

ユーザー企業の立場

生成 AI 活用サービスを利用する組織では、
新たに以下の領域での活躍が期待されるだろう

- **サービスへの入力情報の機密性検討**
  どのような情報をサービスに入力して良いか検討して、自組織内で周知する必要がある
- **サービスからの出力情報の安全性検討**
  出力された情報の正確性・適合性を評価するよう、利用者に教育する必要がある
- **サービスの技術的側面の考慮**
  サービスを導入する (又は、しない) 上でのシステム的な影響を検討する必要がある

ベンダー企業の立場

生成 AI 活用サービスを提供する組織、及び生成 AI の基盤モデルを開発又は提供する組織では...

- **生成 AI サービスの可用性・機密性の維持**
  提供するサービスの安定的運用、入力された情報の適切な管理 (保管、匿名化、廃棄等)
- **プライバシー・ライセンスの管理**
  基盤モデルを生成、又はモデルをチューニングするのに利用する情報の法的適合性の管理
- **生成 AI 固有の特性を悪用した攻撃への対処**
  敵対的 (adversarial) な入力によって、生成 AI に不適切な回答をさせる攻撃への対処

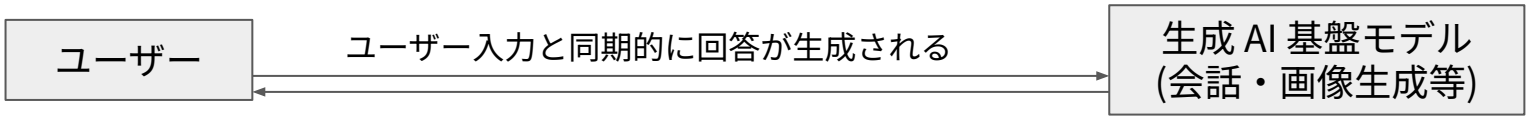
あくまで一例。また、生成 AI を直接利用しない組織でも、無関係ではいられない (後述)

生成 AI 活用サービスの利用シーンに着目すると…

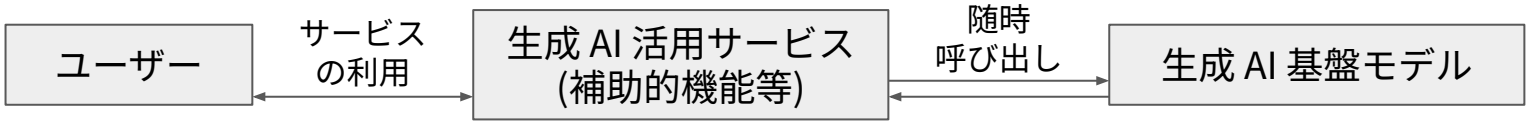
サイバーセキュリティ担当者は、生成 AI 関連サービスに様々な形態があると理解しておく必要がある。これは、求められるスキルにも影響する (次ページ)

サービスの高度化

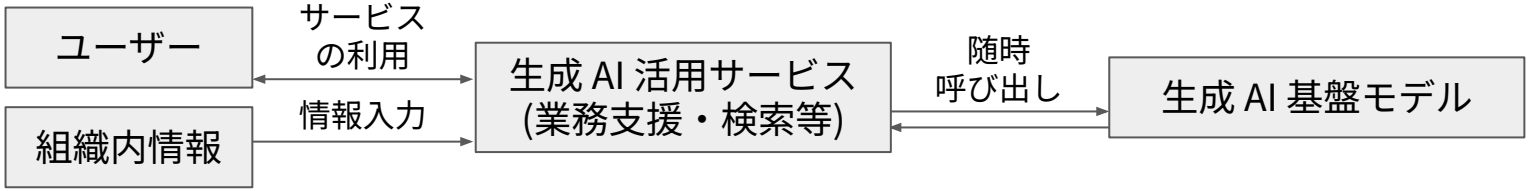
1. ユーザー入力が (ほぼ) 直接的に生成 AI に入力される状況



2. サービスに入力した情報が、加工されるなどして生成 AI に入力される状況



3. RAG (検索拡張生成) のような手法で、ユーザー入力と別に情報が入力される状況



(これ以上に高度な手法も、さらに開発が進んでいる)

生成 AI 活用サービスの利用シーンに着目すると…

以下のようなスキルが新たにサイバーセキュリティ担当者に求められる、と言える

管理的スキル

- 生成 AI 活用サービスがどのように情報を管理するのか理解するスキル
  (既存のベンダー管理業務と共通する。外部委託先としてのリスク評価等)
- サービス利用の利益とリスクを評価し、経営陣に許可又は禁止を提言するスキル
- 生成 AI サービスの利用ポリシーを定め、自組織に周知・研修するスキル
  (どのような方法で、どの情報を入力してよいか、又は、いけないか、を決める)

技術的スキル

- 評価対象の AI サービスがどのような構成なのか理解し、問題を予見するスキル
  (例えば、AI サービスのシステム設計・入力情報の利用方法の推測、自組織に生じるメリットの予測、セキュリティ上の課題の予見、等)
- 許可していない未知の AI 関連サービスが利用されていないか把握するスキル
  (いわゆるシャドウ IT の防止と同様である。CASB の利用等)
- 生成 AI サービスを導入する場合の、システムの設計・構築・運用・保守スキル

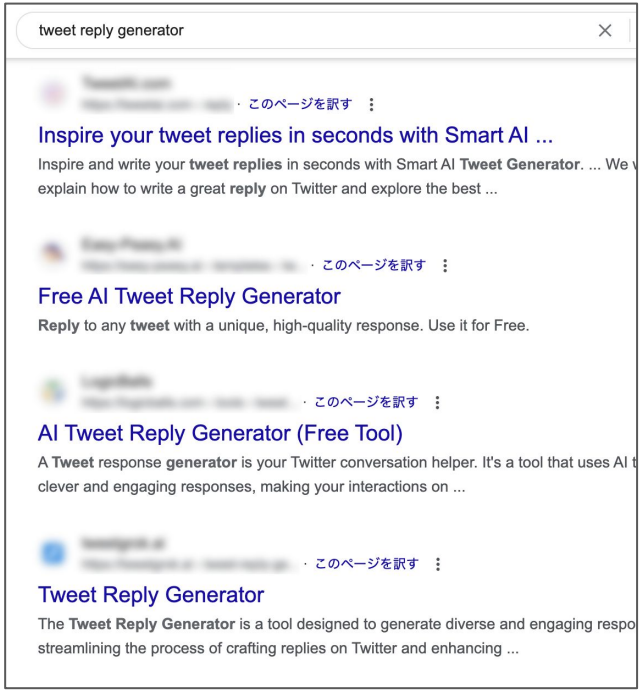
既存スキルの延長

生成 AI を利用しない組織であっても、影響が生じる

自組織で生成 AI を直接利用せずとも、生成 AI を利用する第三者から影響を受けうる

- **生成 AI を使ったスパム行為、架空のユーザー対話**
  X (旧 Twitter) 上では生成 AI による bot が急増。
  虚偽のユーザー対話のための支援ツールが存在 (右図)
- **意図的な虚偽情報やディープ フェイクによる攻撃**
  自組織の役職員に向けた標的型攻撃として、
  巧妙に作られた虚偽情報が送られてくるかもしれない
- **生成 AI を活用したマルウェア作成の自動化**
  攻撃者はビジネスとしてマルウェアを作成しており、
  生成 AI の活用で開発を高速化していると見られる
  (参考: ChatGPTのセキュリティーリスク)

サイバーセキュリティ担当者は、幅広く情報収集すべきであるし、
情報システムの保護に今まで以上に注力すべきである。
周囲には、生成 AI 時代の IT リテラシを啓蒙する必要がある



"tweet reply generator" で検索すると多数の無料ツールが見つかる

(参考) 生成 AI 活用サービスのベンダー向けの検討点

- **学習に利用するデータの保全**
  基盤モデルの作成やファイン チューニング (PEFT 等) に用いるデータセットに細工された情報が混入することで、悪影響が生じる (学習結果の忘却や、不正確な回答など) 可能性がある
- **モデルに対する敵対的な入力 (プロンプト インジェクション等を含む) の防止**
  悪意ある細工された入力は、分類タスクでの誤りやいわゆるジェイルブレイク (指示の逸脱、特定目的 AI の汎化) を惹起する。生成 AI と連動する他のシステムでの誤動作にもつながる
- **モデルからの出力に対するハーネス**
  いわゆるハルシネーション (架空情報の作出) や望ましくない出力を含まないよう、モデルからの出力結果を評価・検証する必要がある
- **学習に用いた情報やモデルそのものの機密性に対する攻撃からの保護** (差分プライバシー等)
  細工した入力によって得られる出力からモデルのパラメータを数理的に分析して不正にモデルを複製する攻撃や、盗んだモデルから学習データを復元する攻撃などが研究されている
- **出力結果の悪用に対する防御**
  生成 AI からの出力が、他所で想定外に転用 (例えば、学生のレポート作成代行から、ディープフェイク等の偽画像など) されないよう、電子透かしを挿入するなどの手法が研究されている

(参考: OWASP Top 10 for Large Language Model Applications)

生成 AI を活用したセキュリティ業務

セキュリティ担当者自らが生成 AI を活用することで、セキュリティ業務を効率化できる。
単純なチャット AI で実現可能な作業がある。既存製品に統合された専用ツール等もある。
例えば、次のようなことが可能:

プログラム作成・書き換え

- 日本語の指示から、管理スクリプトを生成する
- 既存のコードを分析し安全な実装に書き換える

  参考: GitHub Copilot Update: New AI Model That Also Filters Out Security Vulnerabilities

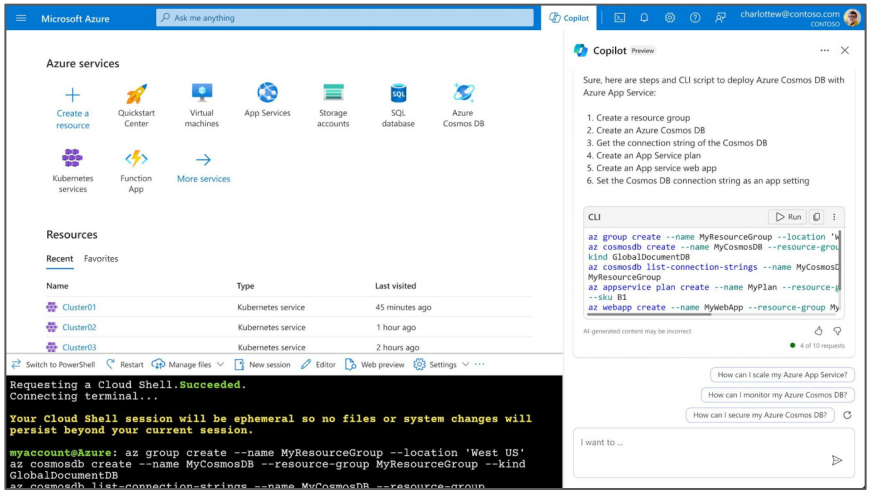
情報の集約・パターン マッチ

- SIEM (ログ基盤) と接続し、社内システムでの不正の兆候を発見・分析する

社内ヘルプデスク

- 社内システムに関する知識をモデルに与え、自動チャットによる IT サポートを提供する

(例) Microsoft Copilot for Azure
クラウド環境の管理画面上で、チャット AI に質問しながら、適切な設定コマンドを確認できる



出典: Simplify IT management with Microsoft Copilot for Azure – save time and get answers fast

生成 AI がセキュリティの人材・スキルに与える影響

- **生成 AI によってセキュリティの仕事がなくなるか？**
  (少なくとも、直ちには) そのようなことはない。
  訓練データで入力された知識と、それを組み合わせた機械学習的な推論を提供してくれるが、セキュリティ業務が置き換えられるにはまだ遠い

- **チャット AI に聞けるなら、セキュリティ技術の知識は不要になるか？**
  適切な質問をし、回答を理解するために、これまで同様に基礎知識は不可欠。
  生成 AI の基盤モデルは、訓練に使用した知識以外を持たないことに注意がいる。
  RAG 等がなければ、特定ドメイン (例えば自組織など)・最新情報については不知

- **セキュリティ業務に従事するのに、経験は不要になるか？**
  生成 AI による支援ツールを使うことで、ある程度は経験不足が補われるだろう。
  むしろ、ツールを使うことで、効率的な習熟につながるかもしれない。
  新しい状況への適応力、知識の吸収力、状況を的確に把握できる思考力等が重要

生成 AI 時代に求められるセキュリティ人材像

自組織・ユーザーが生成 AI を安全に利用できるよう、専門的知見を提供できる

自組織で生成 AI 活用サービスを利用する場合には、その導入・運用が適切に行われるよう、専門家として知見を提供し、ガイドする立場を目指すべきである

- **サイバーセキュリティの基本的素養を、本質から深く理解している**
- **生成 AI の活用方法について情報収集し、システムの特性・利用構成を理解している**
  生成 AI 活用サービスの導入にあたって、既存のセキュリティの知見を活用することは必須。
  既存の知識・能力の重要度は、高まることはあっても、低くなることはない

生成 AI を積極的に利用・評価できるユーザーを、自ら目指している

生成 AI を否定することは困難。チャット形式で利用できる LLM は爆発的に普及しており、抜け穴なく技術的に禁止するほうが難しい。生成 AI を受け入れ、そのリスクを正しく理解するほうが良い

- **自身で生成 AI を活用したサービスを利用し、体験している**
- **生成 AI そのものや、ひいては機械学習・データ処理技術の理解にむけて努力している**
  コンピュータ科学の理解、論理的思考力、言語運用能力等は、さらに重要になる

生成 AI 時代のセキュリティ人材育成に向けた考え方

- **既存のサイバーセキュリティに対する本質的な理解を疎かにさせない**
  セキュリティ全体を俯瞰するための知識体系は今後も身につける必要あり。
  チャット AI などに聞いたとして、回答の正しさを判断・調査できる能力が必要
- **実践的な方針で人材を育てるべきであり、近道はない**
  基礎素養・座学的な知識のみならず、新技術に触れて理解することも重要。
  今は未だ「生成 AI の技術経験 XX年」のような人材は稀少。スタート地点は皆同じ
- **積極的な情報収集を奨励する**
  生成 AI 技術は急発展しており、様々な分野で AI 活用サービスが登場している。
  自組織での生成 AI 活用の可能性・その影響を積極的に検討・評価できるように、
  また組織内外での未知のリスクを予見できるよう、情報収集を奨励すると良い

DX推進スキル標準の他の人材類型との関係

サイバーセキュリティ人材は、生成 AI 利用の後押しをできるよう、他の類型の人材と協調することが望ましい

ソフトウェア エンジニア

サイバーセキュリティ人材は、ソフトウェア開発における知見を有する立場として

- 生成 AI 技術を活用するサービスの開発時に、安全かつ倫理的なサービスを提供できるよう助言する
- 自動生成されたプログラムの安全性等について知見を提供する
  (参考: Can you trust ChatGPT's package recommendations?)
- 生成 AI 技術を利用して行われる新たな形態のサイバー攻撃等に対して、対応策の実装を支援する

等の場面で、エンジニアを補助すべきだろう

ビジネス アーキテクト

生成 AI 技術を利用する際に、セキュリティ上の課題や運用コスト等を提言し、利用可否の意思決定を支援する

デザイナー

サービスのユーザー体験 (UX) において、

- 生成 AI に入力情報が渡されること、
- 表示されている画像・文章が生成 AI によって出力されたものであること、

を的確に伝達できるよう、連携する

データ サイエンティスト

生成 AI 技術を利用するサービスの開発等で利用するデータに対し、

- 処理対象データに対するアクセス管理
- 社内外でのデータ利用者の特定

等において協力することができる